American Journal of Epidemiology
© The Author 2012. Published by Oxford University Press on behalf of the Johns Hopkins Bloomberg School of Public Health. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

Vol. 175, No. 6
DOI: 10.1093/aje/kwr333
Advance Access publication:
February 16, 2012

Original Contribution

On the Relations Between Excess Fraction, Attributable Fraction, and Etiologic Fraction

Etsuji Suzuki*, **Eiji Yamamoto**, and **Toshihide Tsuda**

* Correspondence to Dr. Etsuji Suzuki, Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, 2-5-1 Shikata-cho, Kita-ku, Okayama 700-8558, Japan (e-mail: etsuji-s@cc.okayama-u.ac.jp).

Initially submitted September 17, 2010; accepted for publication August 26, 2011.

It has been noted that there is ambiguity in the expression "attributable fraction," and epidemiologic literature has drawn a distinction between "excess fraction" and "etiologic fraction." These quantities do not necessarily approximate one another, and the etiologic fraction is not generally estimable without strong biologic assumptions. In previous studies, researchers have explained the relations between excess and etiologic fractions in the potential-outcome framework, and few authors have explained the relations between these concepts by showing the correspondence between the potential-outcome model and the sufficient-cause model. In this article, the authors thoroughly clarify the conceptual relations between excess, attributable, and etiologic fractions by explicating the correspondence between these 2 models. In so doing, the authors take into account the potential completion time of each sufficient cause, which contributes to further insight to clarify the 2 types of etiologic fraction, i.e., accelerating etiologic proportion and total etiologic proportion. These 2 measures cannot be distinguished in epidemiologic data, and the differences might be subtle. However, they are closely related to a very fundamental issue of causal inference, that is, how researchers define etiology. Further, the authors clarify the relation between 3 distinct assumptions—positive monotonicity, no preventive action (or sufficient-cause positive monotonicity), and no preventive sequence.

causal inference; monotonicity; potential outcomes; sufficient causes

More than 2 decades ago, Greenland and Robins (1) noted that there is ambiguity in the expression "attributable fraction" and drew a distinction between "excess fraction" and "etiologic fraction." These quantities do not necessarily approximate one another, and the etiologic fraction is not generally estimable without strong biologic assumptions (1). More detailed statistical discussions were addressed in related articles (2, 3), and other papers were meant to be educational for general readers (4–6). In these studies, researchers have explained the relations between excess and etiologic fractions in the potential-outcome framework (7). In some recent articles, investigators have discussed the concept of attributable fractions in the sufficient-component cause framework and have described how redundancy of sufficient causes impacts epidemiologic effect measures (8–12).

In this article, we aim to clarify the relations between excess fractions, attributable fractions, and etiologic fractions in detail by explicating the correspondence between the potential-outcome model and the sufficient-cause model. In so doing, we take into account the potential completion time of each sufficient cause, which further clarifies how researchers should define etiology. Further, as we explain below, in most studies investigators have (sometimes implicitly) made assumptions such as positive monotonicity, no preventive action, and no preventive sequence, which might have resulted in some confusion regarding these concepts. Thus, we also aim to clarify the relation between these assumptions. To avoid technical complications, we do not discuss additional problems that can arise when exposure has multiple levels or when competing risks are being considered.

POTENTIAL OUTCOMES FOR A BINARY EXPOSURE VARIABLE

We let E denote a binary cause of interest ($1 =$ exposed, $0 =$ unexposed) and Y denote a binary outcome of interest ($1 =$ outcome occurred, $0 =$ outcome did not occur). Then, we let Y_{ei} denote the potential outcomes for individual i if, possibly contrary to fact, there had been interventions to set $E = e$ (7). For each individual i, there would thus be 2 possible potential outcomes Y_{1i} and Y_{0i} corresponding to what would have happened to that individual when that person was

Table 1. Correspondence Between Response Types, Risk Status Types, and Sequence Types Under a Binary Exposure and a Binary Outcome[a]

Response Types					Risk Status Types						Sequence Types					
Type	Potential Outcomes		Proportion in Subcohorts[b]		Type	Background Factors			Proportion in Subcohorts[b]		Type	Sequence of Potential Completion Time	d_{1i}	d_{0i}	Proportion in Subcohorts[b]	
	Y_{1i}	Y_{0i}	$E = 1$	$E = 0$		A_1	A_2	A_3	$E = 1$	$E = 0$					$E = 1$	$E = 0$
1	1	1	p_1	q_1	1[c]	1	1	1	r_1	s_1	1	$d_{\phi i} < d_{ei} < d_{\bar{e}i} \leq t_i$	$d_{\phi i}$	$d_{\phi i}$	v_1	w_1
											2	$d_{\phi i} < d_{\bar{e}i} < d_{ei} \leq t_i$	$d_{\phi i}$	$d_{\phi i}$	v_2	w_2
											3	$d_{ei} < d_{\phi i} < d_{\bar{e}i} \leq t_i$	d_{ei}	$d_{\phi i}$	v_3	w_3
											4	$d_{ei} < d_{\bar{e}i} < d_{\phi i} \leq t_i$	d_{ei}	$d_{\bar{e}i}$	v_4	w_4
											5[d]	$d_{\bar{e}i} < d_{\phi i} < d_{ei} \leq t_i$	$d_{\phi i}$	$d_{\bar{e}i}$	v_5	w_5
											6[d]	$d_{\bar{e}i} < d_{ei} < d_{\phi i} \leq t_i$	d_{ei}	$d_{\bar{e}i}$	v_6	w_6
					2	1	1	0	r_2	s_2	7	$d_{\phi i} < d_{ei} \leq t_i < d_{\bar{e}i}$	$d_{\phi i}$	$d_{\phi i}$	v_7	w_7
											8	$d_{ei} < d_{\phi i} \leq t_i < d_{\bar{e}i}$	d_{ei}	$d_{\phi i}$	v_8	w_8
					3[c]	1	0	1	r_3	s_3	9	$d_{\phi i} < d_{\bar{e}i} \leq t_i < d_{ei}$	$d_{\phi i}$	$d_{\phi i}$	v_9	w_9
											10[d]	$d_{\bar{e}i} < d_{\phi i} \leq t_i < d_{ei}$	$d_{\phi i}$	$d_{\bar{e}i}$	v_{10}	w_{10}
					4	1	0	0	r_4	s_4	11	$d_{\phi i} \leq t_i < d_{ei} < d_{\bar{e}i}$	$d_{\phi i}$	$d_{\phi i}$	v_{11}	w_{11}
											12	$d_{\phi i} \leq t_i < d_{\bar{e}i} < d_{ei}$	$d_{\phi i}$	$d_{\phi i}$	v_{12}	w_{12}
					5[c]	0	1	1	r_5	s_5	13	$d_{ei} < d_{\bar{e}i} \leq t_i < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{13}	w_{13}
											14[d]	$d_{\bar{e}i} < d_{ei} \leq t_i < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{14}	w_{14}
2	1	0	p_2	q_2	6	0	1	0	r_6	s_6	15	$d_{ei} \leq t_i < d_{\phi i} < d_{\bar{e}i}$	d_{ei}	$d_{\phi i}$	v_{15}	w_{15}
											16	$d_{ei} \leq t_i < d_{\bar{e}i} < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{16}	w_{16}
3[e]	0	1	p_3	q_3	7[c]	0	0	1	r_7	s_7	17[d]	$d_{\bar{e}i} \leq t_i < d_{\phi i} < d_{ei}$	$d_{\phi i}$	$d_{\bar{e}i}$	v_{17}	w_{17}
											18[d]	$d_{\bar{e}i} \leq t_i < d_{ei} < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{18}	w_{18}
4	0	0	p_4	q_4	8	0	0	0	r_8	s_8	19	$t_i < d_{\phi i} < d_{ei} < d_{\bar{e}i}$	$d_{\phi i}$	$d_{\phi i}$	v_{19}	w_{19}
											20	$t_i < d_{\phi i} < d_{\bar{e}i} < d_{ei}$	$d_{\phi i}$	$d_{\phi i}$	v_{20}	w_{20}
											21	$t_i < d_{ei} < d_{\phi i} < d_{\bar{e}i}$	d_{ei}	$d_{\phi i}$	v_{21}	w_{21}
											22	$t_i < d_{ei} < d_{\bar{e}i} < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{22}	w_{22}
											23[d]	$t_i < d_{\bar{e}i} < d_{\phi i} < d_{ei}$	$d_{\phi i}$	$d_{\bar{e}i}$	v_{23}	w_{23}
											24[d]	$t_i < d_{\bar{e}i} < d_{ei} < d_{\phi i}$	d_{ei}	$d_{\bar{e}i}$	v_{24}	w_{24}

[a] We consider a binary exposure E and a binary outcome Y. We consider 2 potential outcomes, Y_{ei}, for individual i. We consider 3 different types of sufficient causes for outcome Y along with certain binary background factors, as follows: A_1, A_2E, and $A_3\bar{E}$. We let $d_{\phi i}$, d_{ei}, and $d_{\bar{e}i}$ denote the potential completion times of sufficient causes A_1, A_2E, and $A_3\bar{E}$ at which the outcome would occur in individual i, respectively. We also let d_{1i} and d_{0i} denote the potential outcome occurrence time of individual i when exposed ($E = 1$) and unexposed ($E = 0$), respectively. In other words, we denote $d_{1i} = \min\{d_{\phi i}, d_{ei}\}$ and $d_{0i} = \min\{d_{\phi i}, d_{\bar{e}i}\}$. Further, t_i denotes a maximum follow-up time of individual i.

[b] We let p_j and q_j, $j = 1$–4, be proportions of response type j in the exposed and unexposed subcohorts, respectively. Similarly, we define r_j, s_j, v_j, and w_j.

[c] Under the assumption of no preventive action, or sufficient-cause positive monotonicity (i.e., $A_3 = 0$ for $\forall i$), these risk status types are excluded.

[d] Under the assumption of no preventive sequence (i.e., $d_{1i} \leq d_{0i}$ for $\forall i$), these sequence types are excluded.

[e] Under the assumption of (counterfactual) positive monotonicity (i.e., $Y_{0i} \leq Y_{1i}$ for $\forall i$), this response type is excluded.

exposed and unexposed, respectively. As a result, individuals can be classified into 4 (i.e., 2^2) different response types, as enumerated in Table 1 (13). We let p_j and q_j, $j = 1$–4, be proportions of response type j in the exposed and unexposed subcohorts, respectively. In some cases, the effect of the cause E may be in the same direction for all individuals. We say that E has a positive monotonic effect on Y if Y_{ei} is nondecreasing in e for all individuals, that is, $Y_{0i} \leq Y_{1i}$ for $\forall i$ (14, 15), which excludes response type 3.

Throughout this article, we will assume that the consistency assumption is met (16–18), which implies that the observed outcome for individual i is the potential outcome, as a function of intervention, when the intervention is set to the observed exposure.

SUFFICIENT CAUSES

In the sufficient-component cause framework (19), each sufficient cause for the outcome might require the presence of E or the presence of $\bar{E}$ or may not require either, where we let $\bar{E}$ denote the complement of E in the terminology of events. We could thus enumerate 3 different types of sufficient causes for Y along with certain background causes A_k: A_1, A_2E, and $A_3\bar{E}$. Here, A_k denotes a set of all components or factors, other than the presence of E and $\bar{E}$, that may be required for a particular mechanism to operate. For simplicity, we denote the presence of these background causes as $A_k = 1$ and their absence as $A_k = 0$. An individual is at risk of, or susceptible to, sufficient cause k if A_k is present for that person. Note that an individual is of one and only one response type in the potential-outcome framework, whereas an individual may be at risk of none, one, or several sufficient causes. Indeed, we can enumerate 8 (i.e., 2^3) patterns of possible risk status for sufficient causes (Table 1). We let r_j and s_j, $j = 1$–8, be proportions of risk status type j in the exposed and unexposed subcohorts, respectively. Let $\bar{A}_1$ denote the complement of A_1 (i.e., not A_1), let $\bar{A}_2$ denote the complement of A_2, etc. In some cases, we may assume that $\bar{E}$ never acts in a sufficient cause for all individuals—that is, A_3 is present for no individuals (i.e., $A_3 = 0$ for $\forall i$). This assumption is called no preventive action (15), or sufficient-cause positive monotonicity (9) in the sufficient-component cause framework, which excludes risk status types 1, 3, 5, and 7.

In this paper, we further classify these 8 types of risk status to clarify the sequence of the potential completion time of each sufficient cause. We let $d_{\phi i}$, d_{ei}, and $d_{\bar{e}i}$ denote the potential completion times of sufficient causes A_1, A_2E, and $A_3\bar{E}$ at which the outcome would occur in individual i, respectively. We also let d_{1i} and d_{0i} denote the potential outcome occurrence time of individual i when exposed ($E = 1$) and unexposed ($E = 0$), respectively. In other words, we denote $d_{1i} = \min\{d_{\phi i}, d_{ei}\}$ and $d_{0i} = \min\{d_{\phi i}, d_{\bar{e}i}\}$. Further, t_i denotes a maximum follow-up time of individual i. We assume that each potential completion time is different. (If 2 sufficient causes share a component and the shared component is their last acquired component, they are completed simultaneously. This phenomenon is called "overdetermination" as a result of the dependent mechanism (10).) Furthermore, we let $d_{1i} = \infty$ if the outcome would never occur for individual i set $E = 1$, and similarly define $d_{0i} = \infty$. We can thus enumerate 24 (i.e., 4!) sequence types (Table 1). We let v_j and w_j, $j = 1$–24, be proportions of sequence type j in the exposed and unexposed subcohorts, respectively. In some cases, we may assume that d_{1i} is always less than or equal to d_{0i} for all individuals, that is, $d_{1i} \leq d_{0i}$ for $\forall i$ (2). We call this assumption no preventive sequence, which excludes sequence types 5, 6, 10, 14, 17, 18, 23, and 24. Note that the assumption of no preventive sequence is neither a necessary condition nor a sufficient condition for the assumption of no preventive action, and vice versa. On the other hand, both no preventive sequence and no preventive action imply positive monotonicity.

Note that we will not be using the sequence types for discussion of excess fractions or attributable fractions, but we will make use of them for etiologic fractions.

EXCESS FRACTIONS

The excess fraction has been broadly defined as the excess caseload due to exposure (20). Several types of excess fractions can be obtained according to different target populations (1). As we explain below, in this article we propose to distinguish those fractions for which the numerator is included in the population defined by the denominator, and we consistently use "proportion" to refer to these measures (18). When the numerator and the denominator are distinct quantities, neither is included in the other: Such measures are not proportions, and thus we call these measures "caseload."

First, we use the total population as a target. Previous studies have defined the excess fraction as an incidence difference relative to the total incidence under exposure (20, 21). In cohort studies, this measure is equal to an excess risk relative to the exposed risk (20, 21), which can be described as follows:

$$\begin{aligned}\frac{P[Y_1 = 1] - P[Y_0 = 1]}{P[Y_1 = 1]} &= \frac{\sum_e P[Y_1 = 1|\, E = e]P[E = e] - \sum_e P[Y_0 = 1|\, E = e]P[E = e]}{\sum_e P[Y_1 = 1|\, E = e]P[E = e]} \\ &= \frac{\{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)\} - \{\pi(p_1 + p_3) + (1 - \pi)(q_1 + q_3)\}}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)} \\ &= \frac{\pi(p_2 - p_3) + (1 - \pi)(q_2 - q_3)}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)}, \qquad (1)\end{aligned}$$

where π denotes the probability of exposure in the total population, $\pi = P[E = 1]$. As equation 1 shows, the numerator is not included in the denominator, and this measure ranges from $-\infty$ to 1. This measure is therefore not a proportion. In this paper, we call this measure an excess caseload (population), which can be interpreted as an incidence difference (either reduction or increment) when the population was entirely unexposed relative to the incidence when the population was entirely exposed. This is a very general form of excess caseload, and we can describe the excess caseload (exposed), which is defined as an incidence difference when the exposed was unexposed relative to the incidence among the exposed (2). This can be described as follows:

$$\begin{aligned}\frac{P[Y_1 = 1 \mid E = 1] - P[Y_0 = 1 \mid E = 1]}{P[Y_1 = 1|\, E = 1]} &= \frac{(p_1 + p_2) - (p_1 + p_3)}{p_1 + p_2} = \frac{p_2 - p_3}{p_1 + p_2} \\ &= \frac{P[\bar{A}_1 A_2 \bar{A}_3 \mid E = 1] - P[\bar{A}_1 \bar{A}_2 A_3 \mid E = 1]}{P[A_1 \cup A_2|\, E = 1]} \\ &= \frac{r_6 - r_7}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6}. \qquad (2)\end{aligned}$$

This measure can be obtained by substituting 1 for π in the excess caseload (population). Likewise, we can obtain the excess caseload (unexposed) by substituting 0 for π.

To calculate the excess risk as a ‘‘proportion’’ of the total risk under exposure, the numerator should be included in the population defined by the denominator. Algebraically, this means that we need to subtract the joint probability of $Y_1 = 1$ and $Y_0 = 1$ from the marginal probability of $Y_1 = 1$ in the numerator. Therefore, the excess proportion (population) can be defined as follows:

$$\begin{aligned}\frac{P[Y_1 = 1] - P[Y_1 = 1,\ \ Y_0 = 1]}{P[Y_1 = 1]} &= \frac{\sum_e P[Y_1 = 1|\ E = e]P[E = e] - \sum_e P[Y_1 = 1,\ \ Y_0 = 1|\ E = e]P[E = e]}{\sum_e P[Y_1 = 1|\ E = e]P[E = e]} \\ &= \frac{\{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)\} - \{\pi p_1 + (1 - \pi)q_1\}}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)} \\ &= \frac{\pi p_2 + (1 - \pi)q_2}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_2)}. \qquad (3)\end{aligned}$$

This measure can be interpreted as a proportion of cases that would not have occurred when the population was entirely unexposed, relative to cases when the population was entirely exposed. This is a general form of excess proportion, and we can calculate the excess proportion (exposed), which is the proportion of cases among the exposed that would not have occurred when the exposed were entirely unexposed (1, 4), as follows:

$$\begin{aligned}\frac{P[Y_1 = 1 \mid E = 1] - P[Y_1 = 1,\ \ Y_0 = 1 \mid E = 1]}{P[Y_1 = 1|\ E = 1]} &= \frac{(p_1 + p_2) - p_1}{p_1 + p_2} = \frac{p_2}{p_1 + p_2} \\ &= \frac{P[\bar{A}_1 A_2 \bar{A}_3 \mid E = 1]}{P[A_1 \cup A_2|\ E = 1]} \\ &= \frac{r_6}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6}. \qquad (4)\end{aligned}$$

This measure can be obtained by substituting 1 for π in the excess proportion (population). Similarly, we can obtain the excess proportion (unexposed) by substituting 0 for π.

Notably, when we calculate the excess caseload and the excess proportion, we exclusively use the counterfactual risk when exposed and the counterfactual risk when unexposed. In other words, these measures express counterfactual contrasts, or causal contrasts (22).

ATTRIBUTABLE FRACTIONS

Often, epidemiologists would also be interested in the reduction in incidence that would be achieved if the population had been entirely unexposed, compared with its ‘‘current’’ exposure pattern, which has sometimes been defined as the attributable fraction (population) (9, 18, 20, 21). Note that this compares the observed risk with the counterfactual risk. Although this measure is closely related to the excess fraction (population), there are subtle differences between these measures. Indeed, as we explain below, when the exposed are used as a target population, both excess fractions and attributable fractions yield exactly the same results if the consistency assumption is met (16–18). In this paper, we propose to distinguish attributable caseload (population) from attributable proportion (population) in a similar manner. Algebraically, the attributable caseload (population) can be obtained by replacing $P[Y_1 = 1]$ with $P[Y = 1]$ in the excess caseload (population) of equation 1, as follows:

$$\begin{aligned}\frac{P[Y = 1] - P[Y_0 = 1]}{P[Y = 1]} &= \frac{\sum_e P[Y = 1|\ E = e]P[E = e] - \sum_e P[Y_0 = 1|\ E = e]P[E = e]}{\sum_e P[Y = 1|\ E = e]P[E = e]} \\ &= \frac{\{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_3)\} - \{\pi(p_1 + p_3) + (1 - \pi)(q_1 + q_3)\}}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_3)} \\ &= \frac{\pi(p_2 - p_3)}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_3)}. \qquad (5)\end{aligned}$$

This measure can be interpreted as a reduction/increment of observed total cases when the population was entirely unexposed. Notably, the numerator is not included in the denominator, and this measure ranges from $-\infty$ to 1. In most cases, the attributable caseload (population) might be one of the most useful measures in public health if an intervention (e.g., vehicle emission control) were implemented to make everyone in the population unexposed.

If we are rather interested in a proportion of observed cases in the total population that would not have occurred when the population was entirely unexposed, we need the following formula:

$$\begin{aligned}\frac{P[Y=1]-P[Y=1,\;\; Y_0=1]}{P[Y=1]} &= \frac{\sum_e P[Y=1|\; E=e]P[E=e]-\sum_e P[Y=1,\;\; Y_0=1|\; E=e]P[E=e]}{\sum_e P[Y=1|\; E=e]P[E=e]}\\ &= \frac{\{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)\}-\{\pi p_1+(1-\pi)(q_1+q_3)\}}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}\\ &= \frac{\pi p_2}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}.\end{aligned} \qquad (6)$$

The numerator is included in the population in the denominator, and we exclusively use attributable proportion (population), referring to the measure in equation 6. We can obtain the attributable caseload (exposed) and attributable proportion (exposed) by substituting 1 for π in equations 5 and 6, respectively. The practicing clinician would be mainly interested in the attributable caseload (exposed). For example, when a physician advises a patient to stop smoking, he or she is in effect telling the (currently exposed) patient that smoking cessation will reduce (and sometimes increase) the risk of all-cause mortality. When the consistency assumption is met (16–18), the attributable caseload (exposed) yields the same result as the excess caseload (exposed). Likewise, the attributable proportion (exposed) yields the same result as the excess proportion (exposed). We can also obtain the attributable caseload (unexposed) and attributable proportion (unexposed) by substituting 0 for π in equations 5 and 6, respectively. Both of them are, by definition, 0; because there are no exposed people, a program to eliminate the exposure would have no effect.

Table 2 summarizes the algebraic definitions of excess fractions and attributable fractions. Notably, when we can assume the positive monotonic effect of E, response type 3 is excluded (i.e., $p_3 = q_3 = 0$ and $r_7 = s_7 = 0$). Then the excess caseloads are equivalent to the excess proportions, and the attributable caseloads are equivalent to the attributable proportions. This might have caused some confusion in how researchers should define excess fractions and attributable fractions. Table 2 also shows how researchers can calculate each measure when exchangeability (i.e., $Y_e \coprod E$ for all values e) is met (13, 23). Neither excess proportions nor attributable proportions are estimable; if excess caseloads and attributable caseloads yield positive values, they give lower bounds of excess proportions and attributable proportions, respectively. Calculations of their upper bounds are shown in the Web Appendix, which is posted on the *Journal*'s Web site (http://aje.oxfordjournals.org/). Under the assumption of positive monotonicity, however, excess proportions and attributable proportions are equivalent to excess caseloads and attributable caseloads, respectively, both of which are estimable.

ETIOLOGIC FRACTIONS

The etiologic fraction has been broadly defined as the fraction of cases that were ‘‘caused’’ by exposure (1–5, 18, 20, 24, 25). Careful consideration of the 24 sequence types can clarify the 2 alternative definitions of etiologic fractions, and these do not coincide except under the assumption of no preventive sequence.

Some researchers may define ‘‘etiology’’ by taking the time of disease occurrence into consideration; that is, exposure is a contributing factor only if a sufficient cause that contains exposure as a component (A_2E) is ‘‘counterfactually’’ the first sufficient cause to be completed for disease to occur. (Note that some literature has discussed the timing of disease occurrence (1–6).) Thus, with regard to the exposed individuals who become at risk of sufficient causes 1 (A_1), 2 (A_2E), and 3 ($A_3\bar{E}$) during the follow-up period, they define that exposure played an etiologic role in the outcome only if the completion of sufficient cause 2 precedes the counterfactual completion time of both sufficient cause 1 and sufficient cause 3. In this paper, we call this definition ‘‘accelerating etiology.’’ When we are interested in the proportion of accelerating etiologic cases to all of the observed cases (18), we need to obtain the accelerating etiologic proportion (population), which is defined as follows:

$$\begin{aligned}\frac{P[\{d_e\le t\},\;\;\{d_e<d_0\}\;|\;E=1]P[E=1]}{P[Y=1]} &= \frac{P[\{d_e\le t\},\;\;\{d_e<d_\phi\},\;\;\{d_e<d_{\bar{e}}\}\;|\;E=1]P[E=1]}{\sum_e P[Y=1|\; E=e]P[E=e]}\\ &= \frac{\pi(v_3+v_4+v_8+v_{13}+v_{15}+v_{16})}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}.\end{aligned} \qquad (7)$$

Table 2. Excess Fractions, Attributable Fractions, and Etiologic Fractions Under a Binary Exposure and a Binary Outcome[a]

	Algebraic Definition	Description in Terms of Proportion	Calculation[b]
Excess fractions (population)			
Excess caseload (population)	$\frac{P[Y_1=1]-P[Y_0=1]}{P[Y_1=1]}$	$\frac{\pi(p_2-p_3)+(1-\pi)(q_2-q_3)}{\pi(p_1+p_2)+(1-\pi)(q_1+q_2)}$	$\frac{P[Y=1\|E=1]-P[Y=1\|E=0]}{P[Y=1\|E=1]}$
Excess proportion (population)	$\frac{P[Y_1=1]-P[Y_1=1,\ \ Y_0=1]}{P[Y_1=1]}$	$\frac{\pi p_2+(1-\pi)q_2}{\pi(p_1+p_2)+(1-\pi)(q_1+q_2)}$	Not available[c]
Excess fractions (exposed)[d]			
Excess caseload (exposed)	$\frac{P[Y_1=1\|E=1]-P[Y_0=1\|E=1]}{P[Y_1=1\|E=1]}$	$\frac{p_2-p_3}{p_1+p_2}=\frac{r_6-r_7}{r_1+r_2+r_3+r_4+r_5+r_6}$	$\frac{P[Y=1\|E=1]-P[Y=1\|E=0]}{P[Y=1\|E=1]}$
Excess proportion (exposed)	$\frac{P[Y_1=1\|E=1]-P[Y_1=1,Y_0=1\|E=1]}{P[Y_1=1\|E=1]}$	$\frac{p_2}{p_1+p_2}=\frac{r_6}{r_1+r_2+r_3+r_4+r_5+r_6}$	Not available[c]
Attributable fractions (population)			
Attributable caseload (population)	$\frac{P[Y=1]-P[Y_0=1]}{P[Y=1]}$	$\frac{\pi(p_2-p_3)}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}$	$\frac{P[Y=1]-P[Y=1\|E=0]}{P[Y=1]}$
Attributable proportion (population)[e]	$\frac{P[Y=1]-P[Y=1,\ \ Y_0=1]}{P[Y=1]}$	$\frac{\pi p_2}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}$	Not available[f]
Attributable fractions (exposed)[g]			
Attributable caseload (exposed)	$\frac{P[Y=1\|E=1]-P[Y_0=1\|E=1]}{P[Y=1\|E=1]}$	$\frac{p_2-p_3}{p_1+p_2}=\frac{r_6-r_7}{r_1+r_2+r_3+r_4+r_5+r_6}$	$\frac{P[Y=1\|E=1]-P[Y=1\|E=0]}{P[Y=1\|E=1]}$
Attributable proportion (exposed)[e]	$\frac{P[Y=1\|E=1]-P[Y=1,Y_0=1\|E=1]}{P[Y=1\|E=1]}$	$\frac{p_2}{p_1+p_2}=\frac{r_6}{r_1+r_2+r_3+r_4+r_5+r_6}$	Not available[c]
Etiologic fractions (population)			
Accelerating etiologic proportion (population)	$\frac{P[\{d_e\leq t\},\ \ \{d_e<d_0\}\|E=1]P[E=1]}{P[Y=1]}$	$\frac{\pi(v_3+v_4+v_8+v_{13}+v_{15}+v_{16})}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}$	Not available[h]
Total etiologic proportion (population)	$\frac{P[\{d_e\leq t\},\ \ \{d_e<d_\phi\}\|E=1]P[E=1]}{P[Y=1]}$	$\frac{\pi(v_3+v_4+v_6+v_8+v_{13}+v_{14}+v_{15}+v_{16})}{\pi(p_1+p_2)+(1-\pi)(q_1+q_3)}$	Not available[h]
Etiologic fractions (exposed)			
Accelerating etiologic proportion (exposed)	$\frac{P[\{d_e\leq t\},\ \ \{d_e<d_0\}\|E=1]}{P[Y=1\|E=1]}$	$\frac{v_3+v_4+v_8+v_{13}+v_{15}+v_{16}}{p_1+p_2}$	Not available[i]
Total etiologic proportion (exposed)	$\frac{P[\{d_e\leq t\},\ \ \{d_e<d_\phi\}\|E=1]}{P[Y=1\|E=1]}$	$\frac{v_3+v_4+v_6+v_8+v_{13}+v_{14}+v_{15}+v_{16}}{p_1+p_2}$	Not available[i]

[a] We let π denote the probability of exposure in the total population. For the definition of other notations, see Table 1.

[b] We show how to calculate each measure when both consistency and exchangeability are met.

[c] An upper bound is given as $\min\left\{1,\ \frac{P[Y=0|E=0]}{P[Y=1|E=1]}\right\}$, whereas a lower bound is given as $\max\left\{0,\ \frac{P[Y=1|E=1]-P[Y=1|E=0]}{P[Y=1|E=1]}\right\}$.

[d] The excess fractions (exposed) can be obtained by substituting 1 for π of the excess fractions (population). Similarly, the excess fractions (unexposed) can be obtained by substituting 0 for π of the excess fractions (population).

[e] The attributable proportion (population) constitutes a lower bound of the accelerating etiologic proportion (population) as well as the total etiologic proportion (population). Similarly, the attributable proportion (exposed) (or, the excess proportion (exposed)) constitutes a lower bound of the accelerating etiologic proportion (exposed) as well as the total etiologic proportion (exposed).

[f] An upper bound is given as $\min\left\{P[E=1\,|\,Y=1],\ \frac{P[Y=0|E=0]P[E=1]}{P[Y=1]}\right\}$, whereas a lower bound is given as $\max\left\{0,\ \frac{P[Y=1]-P[Y=1|E=0]}{P[Y=1]}\right\}$.

[g] The attributable fractions (exposed) can be obtained by substituting 1 for π of the attributable fractions (population). Similarly, the attributable fractions (unexposed) can be obtained by substituting 0 for π of the excess fractions (population). Under the consistency assumption, the attributable caseload (exposed) and the attributable proportion (exposed) are equivalent to the excess caseload (exposed) and the excess proportion (exposed), respectively.

[h] A trivial upper bound is 1, whereas a lower bound is given as $\max\left\{0,\ \frac{P[Y=1]-P[Y=1|E=0]}{P[Y=1]}\right\}$.

[i] A trivial upper bound is 1, whereas a lower bound is given as $\max\left\{0,\ \frac{P[Y=1|E=1]-P[Y=1|E=0]}{P[Y=1|E=1]}\right\}$.

The accelerating etiologic cases are, at maximum, made up of the 6 sequence types (types 3, 4, 8, 13, 15, and 16), all of which contain individuals with d_{ei} being less than $d_{\phi i}$ and $d_{\bar{e}i}$. In other words, the accelerating etiologic proportion refers to the proportion of the diseased for whom the exposure ''sped up'' the time at which the outcome occurred. Although this measure cannot be inferred from epidemiologic data, a lower bound can be given by the attributable caseload (population) if it yields a positive value.

When we are interested in the proportion of accelerating etiologic cases to the observed exposed cases (2), we need to obtain the accelerating etiologic proportion (exposed), which is defined as follows:

$$\begin{aligned}\frac{P[\{d_e \leq t\},\ \{d_e < d_0\} \mid E = 1]}{P[Y = 1 \mid E = 1]} &= \frac{P[\{d_e \leq t\},\ \{d_e < d_\phi\},\ \{d_e < d_{\bar{e}}\} \mid E = 1]}{P[Y = 1 \mid E = 1]} \\ &= \frac{v_3 + v_4 + v_8 + v_{13} + v_{15} + v_{16}}{\sum_{k=1}^{16} v_k} \\ &= \frac{v_3 + v_4 + v_8 + v_{13} + v_{15} + v_{16}}{p_1 + p_2}. \end{aligned} \tag{8}$$

Note that this can also be obtained by substituting 1 for π in the accelerating etiologic proportion (population). An upper bound of accelerating etiologic proportion (exposed) can be described in terms of the prevalence of background factors of sufficient causes as

$$\frac{P[A_2 \mid E = 1]}{P[A_1 \cup A_2 \mid E = 1]} = \frac{r_1 + r_2 + r_5 + r_6}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6} = \frac{r_1 + r_2 + r_5 + p_2}{p_1 + p_2},$$

whereas its lower bound can be described as

$$\frac{P[\bar{A}_1 A_2 \bar{A}_3 \mid E = 1]}{P[A_1 \cup A_2 \mid E = 1]} = \frac{r_6}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6} = \frac{p_2}{p_1 + p_2}.$$

Note that the lower bound is equal to the excess proportion (exposed) and the attributable proportion (exposed) (1–3, 5, 6, 20), both of which are identified from the data under the assumption of positive monotonicity.

By contrast, some may find it reasonable to define that the exposure caused disease if a sufficient cause that contains exposure as a component is actually the first sufficient cause to be completed (8, 10, 18, 20). We call this definition ''total etiology.'' Note that, compared with the definition of accelerating etiology, this is a slightly broad definition and includes 2 more sequence types (types 6 and 14), in which d_{ei} is longer than $d_{\bar{e}i}$. These 2 sequence types constitute ''nonaccelerating etiology,'' and the total number of etiologic cases is comprised of the accelerating etiologic cases and the nonaccelerating etiologic cases. In other words, the total etiologic proportion is described as the proportion of the diseased for whom the exposure is the ''actual cause of the outcome,'' and the total etiologic proportion (population) is defined as follows:

$$\begin{aligned}\frac{P[\{d_e \leq t\},\ \{d_e < d_\phi\} \mid E = 1]P[E = 1]}{P[Y = 1]} &= \frac{P[\{d_e \leq t\},\ \{d_e < d_\phi\} \mid E = 1]P[E = 1]}{\sum_e P[Y = 1 \mid E = e]P[E = e]} \\ &= \frac{\pi(v_3 + v_4 + v_6 + v_8 + v_{13} + v_{14} + v_{15} + v_{16})}{\pi(p_1 + p_2) + (1 - \pi)(q_1 + q_3)}. \end{aligned} \tag{9}$$

Again, this measure cannot be inferred from epidemiologic data, and the attributable caseload (population) constitutes a lower bound if it yields a positive value.

Similarly, when we are interested in the fraction of total etiologic cases of the observed exposed cases, we need to obtain the total etiologic proportion (exposed), which is defined as follows:

$$\begin{aligned}\frac{P[\{d_e \leq t\},\ \{d_e < d_\phi\} \mid E = 1]}{P[Y = 1 \mid E = 1]} &= \frac{v_3 + v_4 + v_6 + v_8 + v_{13} + v_{14} + v_{15} + v_{16}}{\sum_{k=1}^{16} v_k} \\ &= \frac{v_3 + v_4 + v_6 + v_8 + v_{13} + v_{14} + v_{15} + v_{16}}{p_1 + p_2}. \end{aligned} \tag{10}$$

This can be obtained by substituting 1 for π in the total etiologic proportion (population). An upper bound of total etiologic proportion (exposed) can be described in terms of the prevalence of background factors of sufficient causes as

$$\frac{P[A_2 \mid E=1]}{P[A_1 \cup A_2 \mid E=1]} = \frac{r_1 + r_2 + r_5 + r_6}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6} = \frac{r_1 + r_2 + r_5 + p_2}{p_1 + p_2},$$

whereas its lower bound can be described as

$$\frac{P[\bar{A}_1 A_2 \mid E=1]}{P[A_1 \cup A_2 \mid E=1]} = \frac{r_5 + r_6}{r_1 + r_2 + r_3 + r_4 + r_5 + r_6} = \frac{r_5 + p_2}{p_1 + p_2}.$$

Note that the lower bound is subtly larger than that of accelerating etiologic proportion (exposed). When we make an assumption of no preventive sequence, sequence types 6 and 14 are excluded. Thus, both the accelerating etiologic proportion and the total etiologic proportion yield the same result. Although the differences between these 2 concepts might be subtle, they are related to a very fundamental issue of causal inference—that is, how researchers define etiology—so it would be significant to clarify which measures are used on each occasion.

The above discussion clearly shows that either accelerating etiologic proportions or total etiologic proportions can further exceed attributable proportions. Indeed, etiologic fractions can be 1, or 100%, even though attributable proportions may be very small (1, 4, 6, 20); for example, the accelerating etiologic proportion (exposed) can be 1, even though the attributable proportion (exposed) may be very small when the completion time of sufficient cause 2 (d_{ei}) always precedes the completion times of sufficient causes 1 and 3 ($d_{\phi i}$ and $d_{\bar{e}i}$, respectively). Unfortunately, this condition is not testable with epidemiologic data and rarely has any supporting evidence or genuine plausibility (1, 4, 20). On the other hand, both the attributable proportion (exposed) and the accelerating etiologic proportion (exposed) are equal to 0 when the completion time of either sufficient cause 1 or sufficient cause 3 ($d_{\phi i}$ and $d_{\bar{e}i}$, respectively) precedes that of sufficient cause 2 (d_{ei}). Again, this condition is not testable with epidemiologic data (1, 4, 20). Table 2 summarizes the relation between excess fractions, attributable fractions, and etiologic fractions by showing their algebraic definitions. All of the above measures could also be considered conditional on strata of covariates $\boldsymbol{C} = \boldsymbol{c}$.

Finally, it is worthwhile to mention the relation between etiologic fraction and susceptible proportion. Khoury et al. (26) defined the susceptible proportion as the proportion of (exposed) persons who have all other components of a sufficient cause in which the exposure is a component. Although they apparently omitted response type 3 from their discussion (26), the susceptible proportion can be simply described as a numerator of the upper bound of the (either accelerating or total) etiologic proportion (exposed) in terms of the prevalence of background factors of sufficient causes (i.e., $r_1 + r_2 + r_5 + r_6$). Thus, as Greenland and Robins previously noted (1), the class of the susceptible includes but is potentially larger than the class of (either accelerating or total) etiologic cases. Khoury et al. (26) demonstrated that the maximum and minimum estimates of susceptible proportion can be written as $P[Y = 1|E = 1]$ (i.e., $p_1 + p_2 = r_1 + r_2 + r_3 + r_4 + r_5 + r_6$) and $\{P[Y = 1|E = 1]\} - P[Y_0 = 1|E = 1]\}$ (i.e., $\{(p_1 + p_2) - (p_1 + p_3)\} = p_2 - p_3 = r_6 - r_7$), respectively. However, this study clearly shows that its minimum estimate should be described as $\max\{0, (p_2 - p_3)\}$.

DISCUSSION

We have clarified the conceptual relations between excess fractions, attributable fractions, and etiologic fractions in detail by explicating the correspondence between the potential-outcome model and the sufficient-cause model. As the duality between these 2 models shows, the different approaches of causality provide complementary perspectives and can be employed together to improve causal interpretations (27–29), including the issues of mediation and mechanism (30–33). Further, we have enumerated the correspondence between these 2 models by taking into account the potential completion time of each sufficient cause. The enumeration of the 24 sequence types indeed contributes to further insight to clarify the 2 types of etiologic fraction—that is, accelerating etiologic proportion and total etiologic proportion. Although we cannot show the distinction between these 2 measures in epidemiologic data, this issue is closely related to the definition of causality. The practical limitations of these measures, partly arising from the restriction to a binary cause and the assumptions of no competing risks or no unmeasured confounders, should be addressed in future studies. (For related issues, see Greenland (34) and VanderWeele (9).)

In this article, we have highlighted that the validity of any inference can only benefit from explication and critical scrutiny of the assumptions used to derive the inferences. Even if it is justified to make assumptions such as positive monotonicity, no preventive action, and no preventive sequence, researchers should clearly distinguish these assumptions to apply them to data. Estimation of the public health burden is indeed useful for researchers as well as policy-makers and the public, and it has been encouraged among epidemiologists (21). As we have provided the overview, there are a number of measures that quantify the health burden due to specific risk factors for specific diseases. Thus, epidemiologists should carefully determine and explain which measures are used on each occasion.

ACKNOWLEDGMENTS

Author affiliations: Department of Epidemiology, Graduate School of Medicine, Dentistry and Pharmaceutical Sciences, Okayama University, Okayama, Japan (Etsuji Suzuki); Department of Information Science, Faculty of Informatics, Okayama University of Science, Okayama, Japan (Eiji Yamamoto); and Department of Environmental Epidemiology, Graduate School of Environmental Science, Okayama University, Okayama, Japan (Toshihide Tsuda).

Conflict of interest: none declared.

REFERENCES

1. Greenland S, Robins JM. Conceptual problems in the definition and interpretation of attributable fractions. *Am J Epidemiol*. 1988;128(6):1185–1197.
2. Robins JM, Greenland S. Estimability and estimation of excess and etiologic fractions. *Stat Med*. 1989;8(7):845–859.
3. Robins J, Greenland S. The probability of causation under a stochastic model for individual risk. *Biometrics*. 1989;45(4): 1125–1138.
4. Greenland S. Relation of probability of causation to relative risk and doubling dose: a methodologic error that has become a social problem. *Am J Public Health*. 1999;89(8): 1166–1169.
5. Beyea J, Greenland S. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Phys*. 1999;76(3):269–274.
6. Greenland S, Robins JM. Epidemiology, justice, and the probability of causation. *Jurimetrics*. 2000;40(3):321–340.
7. Little RJ, Rubin DB. Causal effects in clinical and epidemiological studies via potential outcomes: concepts and analytical approaches. *Annu Rev Public Health*. 2000;21: 121–145.
8. Allard R, Boivin JF. Measures of effect based on the sufficient causes model. 1. Risks and rates of disease associated with a single causative agent. *Epidemiology*. 1993;4(1):37–42.
9. VanderWeele TJ. Attributable fractions for sufficient cause interactions. *Int J Biostat*. 2010;6(2):5. (doi:10.2202/1557-4679.1202).
10. Gatto NM, Campbell UB. Redundant causation from a sufficient cause perspective. *Epidemiol Perspect Innov*. 2010;7(1):5. (doi:10.1186/1742-5573-7-5).
11. Hoffmann K, Heidemann C, Weikert C, et al. Estimating the proportion of disease due to classes of sufficient causes. *Am J Epidemiol*. 2006;163(1):76–83.
12. Hoffmann K, Flanders WD. Re: ‘‘Estimating the proportion of disease due to classes of sufficient causes’’ [letter]. *Am J Epidemiol*. 2006;164(12):1254–1255.
13. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*. 1986;15(3): 413–419.
14. VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology*. 2007;18(3):329–339.
15. Greenland S, Lash TL, Rothman KJ. Concepts of interaction. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:71–83.
16. Cole SR, Frangakis CE. The consistency statement in causal inference: a definition or an assumption? *Epidemiology*. 2009; 20(1):3–5.
17. VanderWeele TJ. Concerning the consistency assumption in causal inference. *Epidemiology*. 2009;20(6):880–883.
18. Porta MS, ed. *A Dictionary of Epidemiology*. 5th ed. New York, NY: Oxford University Press; 2008.
19. Rothman KJ. Causes. *Am J Epidemiol*. 1976;104(6):587–592.
20. Greenland S, Rothman KJ, Lash TL. Measures of effect and measures of association. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:51–70.
21. Steenland K, Armstrong B. An overview of methods for calculating the burden of disease due to specific risk factors. *Epidemiology*. 2006;17(5):512–519.
22. Maldonado G, Greenland S. Estimating causal effects. *Int J Epidemiol*. 2002;31(2):422–429.
23. Hernán MA. A definition of causal effect for epidemiological research. *J Epidemiol Community Health*. 2004;58(4):265–271.
24. Rothman KJ, Greenland S, Poole C, et al. Causation and causal inference. In: Rothman KJ, Greenland S, Lash TL, eds. *Modern Epidemiology*. 3rd ed. Philadelphia, PA: Lippincott Williams & Wilkins; 2008:5–31.
25. Boslaugh S, ed. *Encyclopedia of Epidemiology*. Thousand Oaks, CA: Sage Publications; 2008.
26. Khoury MJ, Flanders WD, Greenland S, et al. On the measurement of susceptibility in epidemiologic studies. *Am J Epidemiol*. 1989;129(1):183–190.
27. Flanders WD. On the relationship of sufficient component cause models with potential outcome (counterfactual) models. *Eur J Epidemiol*. 2006;21(12):847–853.
28. VanderWeele TJ, Hernán MA. From counterfactuals to sufficient component causes and vice versa. *Eur J Epidemiol*. 2006; 21(12):855–858.
29. Suzuki E, Yamamoto E, Tsuda T. On the link between sufficient-cause model and potential-outcome model. *Epidemiology*. 2011;22(1):131–132.
30. Hafeman DM. A sufficient cause based approach to the assessment of mediation. *Eur J Epidemiol*. 2008;23(11):711–721.
31. VanderWeele TJ. Mediation and mechanism. *Eur J Epidemiol*. 2009;24(5):217–224.
32. VanderWeele TJ. Subtleties of explanatory language: what is meant by ‘‘mediation’’? *Eur J Epidemiol*. 2011;26(5): 343–346.
33. Suzuki E, Yamamoto E, Tsuda T. Identification of operating mediation and mechanism in the sufficient-component cause framework. *Eur J Epidemiol*. 2011;26(5):347–357.
34. Greenland S. Attributable fractions: bias from broad definition of exposure. *Epidemiology*. 2001;12(5):518–520.