Published in final edited form as:

Ann Epidemiol. 2016 June ; 26(6): 389–394.e2. doi:10.1016/j.annepidem.2016.04.010.

Conditions for valid estimation of causal effects on prevalence in cross-sectional and other studies

W. Dana Flanders, MD, DSc[a,b,*], **Mitchel Klein, PhD**[a,c], and **Maria C. Mirabelli, PhD**[d]

[a]Department of Epidemiology, Rollins School of Public Health, Emory University, Atlanta, GA

[b]Department of Biostatistics and Bioinformatics, Rollins School of Public Health, Emory University, Atlanta, GA

[c]Department of Environmental and Occupational Health, Rollins School of Public Health, Emory University, Atlanta, GA

[d]Air Pollution and Respiratory Health Branch, Division of Environmental Hazards and Health Effects, National Center for Environmental Health, Centers for Disease Control and Prevention, Atlanta, GA

Abstract

Purpose—Causal effects in epidemiology are almost invariably studied by considering disease incidence even when prevalence data are used to estimate the causal effect. For example, if certain conditions are met, a prevalence odds ratio can provide a valid estimate of an incidence rate ratio. Our purpose and main result are conditions that assure causal effects on prevalence can be estimated in cross-sectional studies, even when the prevalence odds ratio does not estimate incidence.

Methods—Using a general causal effect definition in a multivariate counterfactual framework, we define causal contrasts that compare prevalences among survivors from a target population had all been exposed at baseline with that prevalence had all been unexposed. Although prevalence is a measure reflecting a moment in time, we consider the time sequence to study causal effects.

Results—Effects defined using a contrast of counterfactual prevalences can be estimated in an experiment and, with conditions provided, in cross-sectional studies. Proper interpretation of the effect includes recognition that the target is the baseline population, defined at the age or time of exposure.

Conclusions—Prevalences are widely reported, readily available measures for assessing disabilities and disease burden. Effects on prevalence are estimable in cross-sectional studies but only if appropriate conditions hold.

Keywords

Prevalence; Causal effects; Validity; Survey; Cross-sectional studies; Target population

[*]Corresponding author. Department of Epidemiology, Department of Biostatistics and Bioinformatics, Emory University, Rollins School of Public Health, Atlanta, GA 30329. Tel.: 404-7278707; fax: 404-727-8737. wflande@sph.emory.edu (W.D. Flanders).

W.D.F. is a consultant at Biogen Idec. The remaining authors have no conflicts of interest to disclose.

Introduction

A now-common way to define causal effects in epidemiology uses a counterfactual framework [1–5]. Expanding on this approach, Flanders and Klein [6] presented a general definition of causal effects as a contrast of parameters of the distribution of multivariate counterfactual outcomes for the same population under two exposure conditions.

This general approach shows that causal effects can be defined using contrasts of prevalences [6,7]. Nevertheless, prevalences are infrequently used to define or estimate effects. For example in cross-sectional studies, prevalence odds ratios are often not viewed as causal measures unless they are a proxy for an incidence rate ratio [8–10]. Grabovschi et al. [11] seemingly echo this view, stating “The reviewed research studies also have some important methodological limitations related mostly to their reliance on survey data, which could preclude causal interpretation and only measure statistical associations and tendencies.”

Although some have defined [6] and others estimated [7] causal effects using prevalence contrasts, the conditions needed for valid estimation in cross-sectional studies have not yet been discussed. Therefore, our purpose here is to discuss valid estimation of causal effects when interest includes disease prevalence itself rather than just disease occurrence. We review the definition of a causal prevalence difference, provide examples, and discuss interpretation. Our main, new, and novel results are the presentation and discussion of assumptions that, when true, assure that causal effects on prevalence can be validly estimated in cross-sectional studies.

Notation and definitions

We assume exposure (E) is dichotomous and occurs at an early age a_0, if at all. Disease (D) can occur at any age, can resolve, in which case we say D is not present, and can recur in people in whom it had resolved.

Notation

The outcome-vector [$D_{i,a}$, $S_{i,a}$] encodes disease status and survival: disease component $D_{i,a}$ is 1 if individual i is alive with disease and 0 otherwise and survival component $S_{i,a}$ is 1 if individual i is alive and 0 otherwise, both at age a. E_i is 1 if individual i was exposed at age a_o and 0 otherwise. Parentheses denote counterfactual outcomes [1,12]: the vector [$D_{i,a}(e)$, $S_{i,a}(e)$] is the value of [$D_{i,a}$, $S_{i,a}$] if E_i had been set to e at age a_0, for $e = 0,1$. In particular, disease component $D_{i,a}(e)$ is 0 if individual i would have died before age a after setting E_i to e, but other definitions are possible [6,13,14].

Because an individual cannot have been both exposed and unexposed at age a_o, one of the outcome vectors [$D_{i,a}(1)$, $S_{i,a}(1)$] or [$D_{i,a}(0)$, $S_{i,a}(0)$] is counterfactual.

Definitions

Clear effect definitions require several components [2,6,15–17], including specification of the target; relevant ages including those when exposure is to be set and the outcome measured (e.g., follow-up periods); how the exposure will be set to the levels considered; and the contrast (e.g., difference or ratio).

Assuming these components are specified, the effect of E on presence of D at age *a* for individual *i* can be defined as $D_{i,a}(1) - D_{i,a}(0)$, Table 1. Because D encodes being alive with disease at age *a*, this effect is on the joint (composite) outcome having disease and being alive. Possible values are −1, 0, and +1. For example, −1 corresponds to being alive with disease presence D at age *a* if unexposed and either not having disease or not being alive if exposed. This definition differs from the typical one wherein those dying without disease are treated as incomplete observations; however, it is akin to the approach of Fine and Grey [18] wherein those dying without disease are treated as nondiseased, complete observations. Disease could potentially have changed several times between exposure at age a_o and age *a*; focusing on age *a* summarizes net changes. The effect of exposure can depend on disease status at age a_o.

To define a population average effect of E on disease presence at age *a*, we must specify the target population (P_0) at age a_0 when exposure is set. Then, we define the causal prevalence difference (*cPD*) at age *a* for P_0 as the prevalence in P_0 if all had been exposed at age a_0, compared with that prevalence if all had been unexposed. In equation form, *cPD* is

$$cPD=\frac{\sum_{i\in P_0} D_{i,a}(1)}{\sum_{i\in P_0} S_{i,a}(1)} - \frac{\sum_{i\in P_0} D_{i,a}(0)}{\sum_{i\in P_0} S_{i,a}(0)} \quad (1)$$

The vectors $[\sum_{i\in P_0} D_{i,a}(e), \sum_{i\in P_0} S_{i,a}(e), e = 0, 1]$ whose components appear in Equation 1 are parameters (means) of the distributions of counterfactual outcome vectors. If exposure affects survival, that would be part of the causal pathway and reflected in *cPD*. Importantly, the target is P_0, not the subpopulation that survives to age *a*.

Of note, Flanders and Klein [6] previously defined causal effects using prevalence ratios rather than differences. Then, the target population was P_f, the full population at baseline. Now the target population (P_0) coincides with P_f, provided the survey population P_1 consists of all survivors from P_f. Thus, apart from changing from ratios to differences, the previous [6] and present definitions essentially coincide.

Estimation

A natural estimator of the population average causal effect of E on presence of D in population P_0 is the observed difference:

$$\widehat{cPD}=\overline{D}_{1,a_1} - \overline{D}_{0,a_1} \quad (2)$$

where an overbar indicates the average; $\overline{D}_{j,a_1} = \sum_{i\in P_0:\ E_i = j} D_{i,a_1}/N_{j,a_1} = \sum_{i\in P_0:E_i=j} D_{i,a_1}/\sum_{i\in P_0:E_i=j} S_{i,a1}$ is the observed prevalence of D at age $a = a_1$ in exposure group *j* (*j* = 1 if exposed, 0 otherwise); and $N_{j,a1}$ is the observed number alive with exposure equal to *j* at age a_1.

In results, we present assumptions that suffice for this estimator to be unbiased in cross-sectional studies; in Appendix 1, we justify this claim, and in Appendix 2, we discuss why the estimator is unaffected by collider bias from conditioning on survival.

Of note, this estimator is simple and involves directly observable variables. As in the definition, deaths including those from "competing risks" are treated realistically, as part of the causal process affecting disease prevalence.

Results

Our goal and main novel result is to state assumptions sufficient for expression [2] to be a valid estimator of causal prevalence differences in cross-sectional surveys. We first motivate the approach by briefly considering experiments.

Randomized experiments

The effects of exposure on disease presence at a specified time after exposure can be estimated in an experiment. Briefly, one identifies and enrolls subjects, say at age a_0. For simplicity, we focus throughout on a specific age at exposure (a_0), although one could include different age groups and calculate a summary measure or model age patterns. We may optionally measure baseline presence of disease (age a_0) and then expose a random subgroup to E or placebo. We follow the cohort to age a_1 and measure disease presence, assuming no dropouts or loss to follow-up.

Sufficient conditions under which estimator 2 validly estimates causal effects in randomized experiments are exchangeability (Table 1) for disease presence and survival [$\mathrm{D}_{i,a}(e)$, $\mathrm{S}_{i,a}(e)$]; independence between people (stable unit treatment value assumption, SUTVA [19,20]); and counterfactual model consistency [4] (Table 1; e.g., $D_{i,a}(j) = D_{i,a}$, if $E_i = j$). The target is the baseline cohort, and all subjects remain under observation, unless death intervenes. These assumptions describe good randomization (exchangeability), no intersubject interference (SUTVA), complete follow-up, and conceptual clarity (consistency), respectively. They should hold in a well-conducted experiment, possibly apart from SUTVA which can depend on characteristics of the exposure and outcome. Example 1 illustrates how causal effects might be estimated using prevalence contrasts. Additional examples are provided by community intervention trials, often randomized, that frequently use prevalence contrasts to estimate causal effects [21]. Cohort studies involve similar assumptions, with the important caveat that, absent randomization, exchangeability is not necessarily expected (see Appendix 1).

Exchangeability need hold only conditional on measured baseline covariates (C; Table 1). Moreover, the target need not be disease free (at baseline). However, the assumption of exchangeability for nonrandomized studies could be implausible if the prevalence at baseline differed by exposure as noted in the discussion.

Example 1

We illustrate estimation of effects on prevalence in an experiment. Kuller et al. [22] studied the effect of a healthy lifestyle intervention on women's low-density lipoprotein (LDL) cholesterol. Women had similar cholesterol levels at baseline, but 4.5 years after randomization, 27% of the women in the lifestyle intervention group had LDL cholesterol <100 mg/dL ("optimal") compared with 16% in the assessment only group. Many measures and comparisons were additionally reported (e.g., average between-group differences), but results included these prevalences illustrating that prevalences can be relevant and informative.

From their results, the prevalence difference is 27% − 16% = 9%. This contrast is not only descriptive, but with our assumptions that should often be plausible in a randomized experiment (previously mentioned), can also be interpreted as estimating the causal effect of the intervention on prevalence of optimal LDL cholesterol. Because prevalence reflects both incidence and duration after onset [23], effect can be on both—as summarized in the prevalence difference. Follow-up was high (95%) with one death, but if the intervention had affected mortality, then $\widehat{cPD}$ would nevertheless estimate the causal effect on prevalence in the randomization cohort, measured in survivors.

Cross-sectional studies

The assumptions that assure unbiasedness of estimator 2 are less straightforward in a cross-sectional survey. We assume that survey participants are randomly selected from a well-defined population of living people P_1 at age a_1 (no surrogates for the deceased). The presence (or absence) of disease at age a_1 ($D_{i,a1}$) and prior exposure at age $a_0 < a_1$ are accurately assessed.

Perhaps the main challenge is defining the target population needed to clearly define causal effects and for which a survey of population P_1 is expected to yield valid estimates of causal prevalence differences. Because effects require time to occur, the target must have been enumerable at age a_0 before measuring the outcome. We can clearly specify the target if population P_1 consists of all survivors from a larger population P_0 that was alive at the time of potential exposure, age a_0. Population P_0 should be definable by observable, contemporaneous factors. If we can specify P_0, then measurement of disease prevalence in the survey provides just the information needed to estimate cPD and had a cohort study of P_0 been done. In particular, summation over $i \in P_0$ appearing in estimator 2 can be replaced by summation over $i \in P_1$ because summands $D_{\mathrm{i,a}1}$ and $S_{i,a1}$ are 0 for individuals who die between age a_0 and a_1. If the survey involves a 100% sample, prevalences in the sample coincide with those from a cohort study of P_0, and if less than 100%, prevalences are valid estimators of them because we assume exposure-specific prevalences in survey participants represent those in P_1 (Appendix 1).

The other assumptions needed for unbiased estimation coincide with those for experiments and cohort studies. Specifically, we need exchangeability for target population P_0 (Table 1; $D_{i,a_1}(e) \amalg E_i$ and $S_{i,a_1}(e) \amalg E_i$ for $i \in P_0$ and SUTVA). These assumptions mean that comparison of disease status among exposed and unexposed survivors informs what would have happened if exposure had been randomized in population P_0 at age a_0 and the cohort followed to age a_1.

Some examples may help illustrate conceptualization of population P_0 (Table 2). Suppose survey respondents, perhaps like respondents to the National Health and Nutrition Examination Survey or Behavioral Risk Factor Surveillance System, are representative of all (noninstitutionalized) U.S. residents, say at age a_1, and that we focus on exposure at age a_0. We exclude recent immigrants from P_1 because they were not in the resident population at younger ages. Population P_0 is then the population of residents at the younger age a_0, $a_1 - a_0$ years earlier. P_1 should be (or represent) all in P_0 who survive to a_1. Emigration from P_0 is permissible if independent of disease, survival, and exposure.

Figure 1 summarizes causal relationships that, if correct, assure the needed exchangeability assumptions for target population P_0 and representativeness of the surveyed population (P_1) using a directed acyclic graph (DAG). Rules for constructing and interpreting DAGs are reviewed in detail elsewhere [24–26]. The DAG shows that exposure E is (being assumed) independent of other causes of disease (U_0^*), and of other causes (U_0) of disease and survival (S_1). Membership in P_1 depends on survival, not emigrating and other factors, but not directly on E. Participation depends on P_1, other factors U_1, but not directly on E. Under these causal patterns relationships, we expect exchangeability for target population P_0 and representativeness survey population P_1. Figure 2 illustrates situations wherein additional causal effects are present (dotted line) with both S_1 and D_1 affecting membership in P_1. We now expect bias as survey population P_1 may not represent all survivors from target population P_0.

If the exposure-specific prevalences in survey participants do "not" represent those in P_1 (i.e., all P_0-survivors), estimator [2] is likely biased. In this case, however, prevalent disease should be associated with survey participation, or with emigration or loss from the surviving population suggesting a common cause (e.g., C_0 in Fig. 3). If we can control for all such common causes, prevalent disease should then be independent of participation, emigration, and loss (assuming no conditioning on a collider), exposure-specific prevalences in participants should consistently estimate those in survivors, and the conditional estimator should be consistent for the effect in [1].

Example 2

Example 2 illustrates the use of prevalence contrasts for estimating effects in a survey (cross-sectional study). Our goal is to estimate the effect of starting smoking at age 18 years versus never starting or starting later on the prevalence of poor or fair health 20 years later at age 38 years. Self-rated health status is of interest partly because it consistently and strongly predicts subsequent mortality, even after control for multiple other health-status indicators [27]. To estimate this effect, we use data from the National Health and Nutrition Examination Survey for the years 2007–2010. To increase sample size, we include respondents 35–39 years old when surveyed (instead of just 38 years). The exposed group is smokers who started regular smoking between ages 18 and 21 years, using a wider age interval to increase sample size, and the unexposed are all who never started or started after age 21 years. The outcome is self-reported poor or fair health at interview (age, 35–39 years). After adjusting for gender and race, the prevalence odds ratio was 1.7, indicating a

70% estimated higher prevalence odds of poorer health among those who started regular smoking at ages 18–21 years, compared with never or later starting smokers.

The baseline population P_0 consisted of U.S. residents who were 18–21 years old about 20 years before interview. If the exposed and unexposed were exchangeable conditional on controlled covariates, and participants were representative of all U.S. residents aged 38 ± 2 years during this time period—the prevalence odds ratio should consistently estimate the effect of taking up regular smoking at about age 18 years on having self-reported poor health 20 years later. Some U.S. residents died between ages 18 and 38 years possibly because of smoking. Because our interest is in the effect of smoking on prevalence, these deaths do not represent bias because of competing risks but rather are part of the defined effect of smoking [6] on subsequent disease prevalence among survivors.

Discussion

When epidemiologists consider disease causation they almost invariably consider it in terms of disease onset (i.e., incidence). Rothman et al. developed causal concepts as follows ([5], p.6): “To begin, we need to define cause. One definition is the cause of a specific disease occurrence is an antecedent event, condition, or characteristic that was necessary for the occurrence of the disease at the moment it occurred, given that other conditions are fixed.” Although their interest in that definition is on disease onset, for effects on disease prevalence, we can similarly consider an antecedent event, condition, or characteristic that was necessary for an individual having disease at a particular point in time, given that other conditions are fixed. Effects on survival, disease onset, or disease duration in this context are part of the causal pathway.

A cohort study is often viewed as a natural design to estimate disease incidence, and a cross-sectional study as a natural design to estimate prevalence as prevalence is a measure reflecting a moment in time. However, to study causal effects, the time sequence must be considered. So, to study causal effects on prevalence, a cohort study would be a natural design. Our assumptions provide conditions wherein observations from a cross-sectional study provide information that adequately approximates information from a cohort study for estimating effects on prevalence.

We have defined causal contrasts that compare the prevalence among survivors from the target population had all in the target been exposed at baseline with that prevalence had they been unexposed. The definition requires specification of the target population, exposure, ages, and other factors [2,6,28]. The assumptions needed for valid estimation are strong and require critical review to assess their validity. Each of these issues merits separate discussion.

A key assumption is that the baseline, target population be clearly defined and potentially enumerable. Although explicitly defined in experiments and cohort studies, in cross-sectional studies, this baseline population may require conceptualization as an earlier “parent” population defined so that the population surveyed would consist of all (represent) survivors from the parent population. Identification of a parent population with the needed characteristics creates a situation wherein observations from a cross-sectional study can reproduce those that would have been obtained if a cohort study of that population had been done. Many surveys will not permit clear delineation of the parent population; if not, associated causal-effect definitions may be unsatisfactory. Other assumptions for validity of the observed PD as an estimator of the causal PD in cross-sectional studies are equally important. In particular, exchangeability must be evaluated and, as in cohort studies, can be suspect. It is not expected to hold if confounding is present, perhaps due to causes of disease that are also associated with exposure. If the target is not disease free at baseline, exchangeability can also be suspect if the prevalence at baseline differs between the exposed and unexposed subgroups. In cross-sectional studies involving prevalence contrasts, exchangeability can be threatened by factors that affect either disease duration or risk and are associated with exposure. As with other observational studies, some assumptions are unverifiable, and sensitivity analyses may be useful.

The ages at exposure and at disease measurement must be clearly specified. These ages are critical for several reasons. First, age is a potential confounder, for example, if it affected prevalence and was associated with exposure. Second, age at exposure could be an effect measure modifier. Third, the age and time intervals between exposure (or nonexposure) and outcome measurement can also affect prevalence.

Ideally, exposure would have occurred, if at all, at age a_0, the age of the target population at baseline, analogous to recommendations that follow-up in cohort studies begins at or before exposure [29–31]. To illustrate how we might define exposure and the target in practice, suppose our goal is to estimate the effect of hormone replacement therapy (HRT) on cardiovascular disease (CVD) prevalence. A population-based survey of 60 year olds is available that included questions about age at starting HRT. We could define "exposure" as having started HRT by a specific age, say 50 years (10 years before the survey) and "nonexposure" as not having started by that age (including never starting). P_0 should consist of people who were 50 years old 10 years before the survey and should be defined so that P_1 consists of all survivors from P_0. P_0 may be easier to define if the survey is population-based (e.g., all 60-year-old, U.S. residents in 2010 excluding recent immigrants), so that it might be defined as the corresponding population, 10 years earlier (e.g., all 50-year-old, U.S. residents in 2000). See Appendix 2 for additional discussion. If the timing and ages are not specified, then the effect may not be clearly defined, and confusion can ensue. Importantly, similar issues can arise in cohort studies. A possible example concerns side effects of HRT, thought by many to be protective for CVD. After randomized trials showed HRT to increase rather than decrease CVD risk, reanalyses of one observational study suggested that better control for time since initiation and confounding could lead to better agreement with the randomized trial. Thus, clear specification of age at, and time since, exposure are important in cohort as well as cross-sectional studies [29,31] with exposure, if it occurs, being at or around the start of follow-up. Here, we restrict to specific ages for simplicity, but a stratified approach with calculation of summary measures or model-based estimation would directly extend our results.

A comparison of prevalences may seem wanting as a causal measure because prevalence is affected by disease incidence and duration (see Appendix 2) [23]. A harmful exposure could increase the prevalence by increasing the rate of disease onset or decrease the prevalence by increasing the case fatality rate. Because the causal prevalence difference is a summary effect, use of additional contrasts such as differences in risk or duration can be helpful, sometimes vital. Nevertheless, prevalence itself is a widely reported and readily available epidemiologic measure for assessing disabilities, disease burden, and frequency, particularly for chronic, incurable diseases with long duration and unclear timing of onset [23]. Causal inference from prevalence has usually been considered as a proxy for incidence. Our purpose here and main novel result is to provide conditions for validly estimating causal effects in cross-sectional studies. However, we also discuss interpretation and related conceptual issues as the use of prevalence contrasts for defining and estimating causal effects is uncommon and involves relatively new considerations.

Acknowledgments

This publication was supported by US EPA grant R834799. This publication's contents are solely the responsibility of the grantee and do not necessarily represent the official views of the US EPA or CDC. Furthermore, US EPA does not endorse the purchase of any commercial products or services mentioned in the publication.

References

1. Rubin DB. Estimating causal effects of treatments in randomized and non-randomized studies. J Educ Psychol. 1974; 66(5):688–701.
2. Hernán MA. A definition of causal effect for epidemiology. J Epidemiol Community Health. 2004; 58:265–271. [PubMed: 15026432]
3. MacMahon, B.; Pugh, T. Causes and entities of disease. In: Clark, DW.; MacMahon, B., editors. Preventive Medicine. Boston: Little Brown, Boston; 1967. p. 11-18.
4. Hernan MA, Robins J. Causal inference [electron]. 2013 [Accessed December 8, 2013] Available from: http://www.hsph.harvard.edu/miguel-hernan/causal-inference-book.
5. Rothman, KJ.; Greenland, S.; Lash, TL. Modern epidemiology. 3rd. Philadelphia: Wolters Kluwer Health/Lippincott Williams & Wilkins; 2008.
6. Flanders WD, Klein M. A general, multivariate definition of causal effects in epidemiology. Epidemiology. 2015; 26(4):481–489. [PubMed: 25946227]
7. Robins JM, Greenland S, Hu FC. Estimation of the causal role of a time-varying exposure on the marginal mean of a repeated binary outcome. J Am Stat Assoc. 1999; 94:687–700.
8. Lee J, Chia K. Estimation of prevalence rate ratios for cross sectional data: an example in occupational epidemiology. Br J Ind Med. 1993; 50(9):861. [PubMed: 8398881]
9. Thompson ML, Myers J, Kriebel D. Prevalence odds ratio or prevalence ratio in the analysis of cross sectional data: what is to be done? Occup Environ Med. 1998; 55(4):272–277. [PubMed: 9624282]
10. Reichenheim ME, Coutinho ES. Measures and models for causal inference in cross-sectional studies: arguments for the appropriateness of the prevalence odds ratio and related logistic regression. BMC Med Res Methodol. 2010; 10:66. [PubMed: 20633293]
11. Grabovschi C, Loignon C, Fortin M. Mapping the concept of vulnerability related to health care disparities: a scoping review. BMC Health Serv Res. 2013; 13(1):94. [PubMed: 23496838]
12. Flanders WD, Eldridge RC. Summary of relationships between exchange-ability, biasing paths and bias. Eur J Epidemiol. 2015; 30:1089–1099. [PubMed: 24894825]
13. Rubin DB. Causal inference through potential outcomes and principal stratification: application to studies with“ censoring” due to death. Stat Sci. 2006; 21:299–309.
14. Zhang JL, Rubin DB. Estimation of causal effects via principal stratification when some outcomes are truncated by"death". J Educ Behav Stat. 2003; 28(4):353–368.
15. Maldonado G, Greenland S. Estimating causal effects. Int J Epidemiol. 2002; 31:422–429. [PubMed: 11980807]

16. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. Epidemiology. 2011; 22(3):368. [PubMed: 21399502]

17. Hernán MA. Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? Am J Epidemiol. 2005; 162(7):618–620. [PubMed: 16120710]

18. Fine JP, Grey RJ. A proportional hazards model for the subdistribution of a competing risk. J Am Stat Assoc. 1999; 94:496–509.

19. Rubin DB. Direct and indirect causal effects via potential outcomes. Scand J Stat. 2004; 31(2): 161–170.

20. Rubin DB. Comment: Neyman (1923) and causal inference in experiments and observational studies. Stat Sci. 1990; 5(4):472–480.

21. Sorensen G, Emmons K, Hunt MK, Johnston D. Implications of the results of community intervention trials. Annu Rev Public Health. 1998; 19(1):379–416. [PubMed: 9611625]

22. Kuller LH, Simkin-Silverman LR, Wing RR, Meilahn EN, Ives DG. Women's Healthy Lifestyle Project: a randomized clinical trial results at 54 months. Circulation. 2001; 103(1):32–37. [PubMed: 11136682]

23. Rothman, KJ. Modern epidemiology. Boston: Little, Brown and Co; 1986.

24. Greenland S, Pearl J, Robins J. Causal diagrams for epidemiologic research. Epidemiology. 1999; 10:37–48. [PubMed: 9888278]

25. Glymour, MM.; Greenland, S. Causal diagrams. In: Rothman, KJ.; Greenland, S.; Lash, TL., editors. Modern epidemiology. 3rd. Philadelphia: Lippincott, Williams & Wilkins; 2008. p. 183-209.

26. Pearl, J. Causality. 2nd. Cambridge: Cambridge University Press; 2009.

27. Idler EL, Benyamini Y. Self-rated health and mortality: a review of twenty-seven community studies. J Health Soc Behav. 1997; 38:21–37. [PubMed: 9097506]

28. Hernán MA, Robins JM. Estimating causal effects from epidemiological data. J Epidemiol Community Health. 2006; 60(7):578–586. [PubMed: 16790829]

29. Flanders WD, Klein M. Properties of 2 counterfactual effect definitions of a point exposure. Epidemiology. 2007; 18(4):453–460. [PubMed: 17473709]

30. Ray WA. Evaluating medication effects outside of clinical trials: new-user designs. Am J Epidemiol. 2003; 158(9):915–920. [PubMed: 14585769]

31. Prentice RL, Langer R, Stefanick ML, et al. Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the Women's Health Initiative clinical trial. Am J Epidemiol. 2005; 162(5):404–414. [PubMed: 16033876]

32. Bickel, PJ.; Klaassen, CAJ.; Ritov, Y.; Wellner, JA. Efficient and adaptive estimation of semiparametric models. New York: Springer-Verlag; 1993.

33. Flanders WD, Eldridge RC, McClellan W. A nearly unavoidable mechanism for collider bias with index-event studies. Epidemiology. 2014; 25(5):762–764. [PubMed: 25051308]

34. Hernán MA, Hernández-Díaz S, Robins JM. A structural approach to selection bias. Epidemiology. 2004; 15(5):615–625. [PubMed: 15308962]

35. Glymour MM, Weuve J, Berkman LF, Kawachi I, Robins JM. When is baseline adjustment useful in analyses of change? An example with education and cognitive change. Am J Epidemiol. 2005; 162(3):267–278. [PubMed: 15987729]

Appendix 1

Experiments

We argue that the estimator (expression 2) is unbiased under our assumptions for experiments. By completeness of follow-up, $N_{j,a_1} = \sum_{i \in P_0 : E_i = j} S_{i,a_1}$ for $j = 0$ or 1, where $S_{i,a1}=1$ if subject i is alive at age a_1 and 0 otherwise. By counterfactual model consistency for both $D_{i,a1}(e)$ and $S_{i,a_1}(e)$, $\sum_{i \in P_0 : E_i = j} D_{i,a_1} / \sum_{i \in P_0 : E_i = j} S_{i,a_1} = \sum_{i \in P_0 : E_i = j} D_{i,a_1}(j)/$

$\sum_{i \in P_0 : E_i = j} S_{i,a_1}$ (j). By exchangeability, E$[\sum_{i \in P_0 : E_i = j} D_{i,a_1}$ $(j)]/E[\sum_{i \in P_0 : E_i = j} S_{i,a_1}$ $(j)] = \sum_{i \in P_0} D_{i,a_1}$ $(j)/\sum_{i \in P_0} S_{i,a_1}$ (j). Slutsky's theorem [32] now implies that expression [2] is unbiased (technically, consistent) for the causal effect (Equation 1).

Cohort studies

To estimate the cPD in a cohort study, we select a cohort P_0, some of whom were exposed at baseline (age a_0), others not. The assumptions needed for unbiasedness of estimator 2 are the same as those for an experiment. However, exchangeability, which should hold in an experiment with good randomization, needs to be critically evaluated. In particular, we must verify that collider bias [4,5,33,34], if any, induced by cohort selection at baseline is negligible. We follow the cohort to age a_1 and assess disease presence.

Our claim, that estimator [2] is unbiased given our assumptions, follows from the preceding arguments as the design and assumptions closely parallel those for an experiment. The key difference for cohort studies is that the assumption of exchangeability is not expected to hold by design, at least absent restriction, stratification or adjustment, and must be evaluated with particular care using all available information.

Cross-sectional studies

Finally, we argue that estimator 2 is consistent under our assumptions for cross-sectional studies. By assumption, the sample is representative of population P_1, so $\widehat{cPD} \approx (\sum_{i \in P_1 : E_i = 1} D_{i,a} / N_{1,a_1} - \sum_{i \in P_1 : E_i = 0} D_{i,a} / N_{0,a_1})$, where the summation is over subjects in population P_1. Also, by assumption, a larger, enumerable population P_0 exists such that P_1 consists of all surviving members of P_0. Because $D_{i,a}$ and $S_{i,a}$ are both 0 for the deceased, we have in expectation, $\widehat{cPD} \approx (\sum_{i \in P_0 : E_i = 1} D_{i,a} / \sum_{i \in P_0 : E_i = 1} S_{i,a} - \sum_{i \in P_0 : E_i = 0} D_{i,a} / \sum_{i \in P_0 : E_i = 0} S_{i,a})$.

This last expression is the same as estimator 2 for the target cohort P_0 if it was the baseline population in a cohort study followed from age a_0 to age a_1. But estimator 2 is unbiased by our assumptions and arguments mentioned previously for the cohort P_0.

At times, an alternative estimator that accounts for baseline prevalence may be unbiased even if estimator 2 is biased.

$$\widehat{cPD} = (\overline{D}_{1,a_1} - \overline{D}_{0,a_1}) - (\overline{D}_{1,a_0} - \overline{D}_{0,a_0}) \quad (3)$$

For example, the prevalence at baseline may differ between exposed and unexposed because of factors associated with exposure but unassociated with changes in disease thereafter. Measurement error can affect the decision to further account or adjust for baseline disease status as discussed by Glymour et al. [35].

Appendix 2

Alternative approach to defining prevalence effects

Here we consider an approach to defining prevalence effects that provides more details, still rooted in the general causal-effect definition which contrasts parameters of the multivariate counterfactual-outcome distributions presented by Flanders and Klein [1]. The idea is to separate and incorporate the three components of prevalence: disease onset, disease duration, and survival. $\mathrm{D}_{i,a}$ defined in the main text can be viewed as an infinite-dimensional vector, $\mathbf{D}_{\boldsymbol{i}}$ that traces the disease and survival status of individual *i* over each moment of the period of interest. The a^{th} component of $\mathbf{D}_{\boldsymbol{i}} = \mathrm{D}_{i,a}$ encodes disease presence, as defined in the main text, for each age $a \geq 0$. Similarly, the a^{th} components of $\mathbf{D}_{\boldsymbol{i}}(e)$, $\mathbf{S}_{\boldsymbol{i}}(e)$ and $\mathbf{S}_{\boldsymbol{i}}$ are $\mathrm{D}_{i,a}(e)$, $\mathrm{S}_{i,a}(e)$ and $\mathrm{S}_{i,a}$ respectively. Using $\mathbf{D}_{\boldsymbol{i}}(e)$, a matrix of counterfactual outcomes consisting of three vectors can be derived:

$$\mathbf{T}_i(e)=\left(t_{i,1}(e), t_{i,2}(e), \ldots, t_{i,n_i(e)-1}(e), t^{\delta}_{i,n_i(e)}(e)\right)$$

$$\mathbf{U}_i(e)=\left(u_{i,1}(e), u_{i,2}(e), \ldots, u_{i,n_i(e)-1}(e), u^{\delta}_{i,n_i(e)}(e)\right)$$

$$\mathbf{M}_i(e)=\left(m^{\delta}_i(e)\right)$$

where $t_{i,k}(e)$ represents the time from baseline (time 0 or age a_0) to the k^{th} disease episode of subject *i*, and $n_i(e)$ is the total number of his/her episodes during the observation period, if E_i were set to *e* at baseline. If subject *i* has disease at baseline then $t_{i,\,1}(e)$ is zero. $u_{i,k}(e)$ represents the duration of the k^{th} episode of subject *i*, if E_i were *e*. If subject *i* has disease at baseline then $u_{i,\,1}(e)$ is the duration of the first episode from baseline. $m^{\delta}_i(e)$ represents survival time from baseline. The superscript δ (in $t^{\delta}_{i,n_i(e)}(e), u^{\delta}_{i,n_i(e)}(e)$ and $m^{\delta}_i(e)$ is 1 if information is censored and 0 otherwise.

This formulation includes details about disease onset and duration for potentially multiple disease episodes over time. With it, causal effects of exposure on prevalence can be traced over time since baseline. Additionally, this information allows consideration of the causal effect of exposure on disease onset, survival, cumulative disease duration, average duration per episode, proportion of time with the disease or condition and the number of episodes.

Comparison of causal prevalence differences (Equation 1) with causal conditional risk differences

One of the examples used by Flanders and Klein to illustrate their general, multivariate definition of causal effects was the causal conditional risk difference (cCRD) [1]. Here we compare the cCRD with the causal Prevalence Difference (cPD) given by Equation (1). The cCRD for the risk of an outcome during a risk period for the target P_0, conditional on survival to age *a*, can be defined for the target P_0 by:

$$\text{cCRD}=\frac{\sum_{i\in P_0} I_{i,a}(1)}{\sum_{i\in P_0} S_{i,a}(1)}-\frac{\sum_{i\in P_0} I_{i,a}(0)}{\sum_{i\in P_0} S_{i,a}(0)} \quad (1A)$$

where $I_{i,a}(e)$ is 1 if the outcome of interest occurs between age *a* and before the end of the risk period and as 0 otherwise. This expression is like Expression 1 of the main text, but $I_{i,a}(e)$ replaces $D_{i,a}(e)$ and is defined differently. The definition of cPD is similar, but involves presence of disease at age *a* (reflected in $D_{i,a}(e)$), rather than occurrence of disease in the risk period starting at age *a* (reflected in $I_{i,a}(e)$).

The causal prevalence difference estimator (expression 2) and potential collider bias

Since the denominator of Estimator 2 equals the number in the baseline population (target P_0) who survive to age a_1, one could, and a reviewer did, ask if the estimator in Equation 2 might be affected by collider bias due to conditioning on survival to age a_1. A theoretical justification that the estimator in Equation 2 is consistent is outlined in Appendix 1. Here, we provide alternative, less technical arguments. (Briefly, that justification uses the assumptions in the main text including exchangeability in the target P_0 and SUTVA to show that $(\sum_{i\in P_0:E_i=e} D_{i,a}/N_{P0,e}, \sum_{i\in P_0:E_i=e} S_{i,a}/N_{P0,e})$ is an unbiased estimator of the population-average, multivariate effect of E on outcome vector (D_i, S_i), where $N_{P0,e}$ is the number in P_0 with E = e. Slutsky's theorem then shows consistency for the ratio contrasts–the prevalence difference.) Collider bias for one target (e.g., P_1) but not another (e.g. P_0), is also discussed elsewhere [1].

First, we emphasize that the target population is selected at baseline (time 0), and the exposure is independent of risk factors for disease and survival in this population (exchangeability assumption, perhaps conditional on common causes). Thus, selection (collider bias) from selection of the target is not an issue, by assumption. Furthermore, the final estimator (equation 2) is merely an algebraic manipulation of the multivariate estimator $(\sum_{i\in P_0:E_i=e} D_{i,a}/N_{P0,e}, \sum_{i\in P_0:E_i=e} S_{i,a}/N_{P0,e})$. But this multivariate estimator involves no exclusions, stratification, or control (except for stratification by exposure which is exchangeable), and so involves no collider bias.

Second, we note reassuringly that the cPD (equation 1) is directly estimable in a well-conducted experiment (using estimator 2), in which exposure is randomized at baseline in the target P_0. The causal prevalence difference addresses the question – "What is the population average effect of exposure on the target population, as measured by disease prevalences at age a_1?" Of course other question can and typically should be asked, such as, "What is the population average effect of exposure on the target population, as measured by survival at age a_1?" or "What is the population average effect of exposure on the target population, as measured by disease incidence through age a_1?" Although other questions exist, by randomizing exposure at baseline, following the exposed and unexposed groups to age a1, and then accurately measuring disease presence and contrasting the prevalences, one can consistently estimate the effect of exposure on prevalence (via estimator 2). Thus, the defined effect (expression 1) is directly observable using simple, well-defined procedures that end by using estimator 2 in a randomized experiment.

Third, we note that effects of exposure, if any, on death before age a1, are part of the defined effect on prevalence (Expression 1), and appropriately reflected in defining the effect and calculating the measure that estimates it. For example, one way in which exposure could cause a reduction in disease prevalence at age a_1 would be to differentially reduce survival among those who had developed disease. Such a prevalence reduction would be an expected part of an effect on prevalence and correctly estimated, in a randomized experiment or other study under our assumptions. Through use of multivariate outcomes and effects, as described above, a more complete characterization of the exposure's impacts can be obtained.

Supplemental Example 1

Supplemental example 1 illustrates use of prevalence contrasts for estimating effects in a cohort study. To estimate effects of early-life factors on sedentary lifestyle in adolescents, Hallal et al. [2] conducted a cohort study of all children born in-hospital, during 1993 in Pelotas, Brazil. They found a sedentary-lifestyle prevalence of 53.5% among 10–12 year olds whose mother had low education, compared to 63.2% among those whose mother had high education. The prevalence difference (9.7%), if exchangeability and our other assumptions are adequately approximated, is interpretable as the effect of maternal education on prevalence of sedentary lifestyle in adolescents and illustrates use and estimation of prevalence contrasts in a cohort study.

Additional considerations in defining exposures for potentially-ongoing exposures

To further illustrate issues that can arise in defining exposure contrasts and the target P_0, consider the effects of starting alcohol use at a young age, say age 15, on prevalence of hepatic disease at age 35. We might define exposure as having started regular heavy alcohol use by age 15, and for comparison an "unexposed" group as those who had not started regular, heavy drinking by age 15. The resulting contrast, just as it would be in a cohort study, actually compares the effect of starting alcohol use early (age 15) versus later or never. The presence of people who later became a regular heavy drinker would, just as in a cohort study, reduce the expected effects of heavy drinking–compared to a completely unexposed population of, say, never drinkers. But, even in a cohort study–accounting for changes in exposure might require G-computation or related method [3] – if those changes reflect time-varying confounding. If a population-based survey of 35 year olds is available, P_0 should be defined, if possible, as those who were 15 years old about 20 years before the survey, in a way that the survey population represents all survivors from P_0.

References

1. Flanders WD, Klein M. A general, multivariate definition of causal effects in epidemiology. Epidemiology. 2015; 26(4):481–489. [PubMed: 25946227]

2. Hallal PC, Wells JC, Reichert FF, Anselmi L, Victora CG. Early determinants of physical activity in adolescence: prospective birth cohort study. BMJ. 2006; 332(7548):1002–1007. [PubMed: 16601016]

3. Robins JM, Blevins D, Ritter G, Wulfsohn M. G-estimation of the effect of prophylaxis therapy for Pneumocystis carinii pneumonia on the survival of AIDS patients. Epidemiology. 1992; 3:319–336. [PubMed: 1637895]

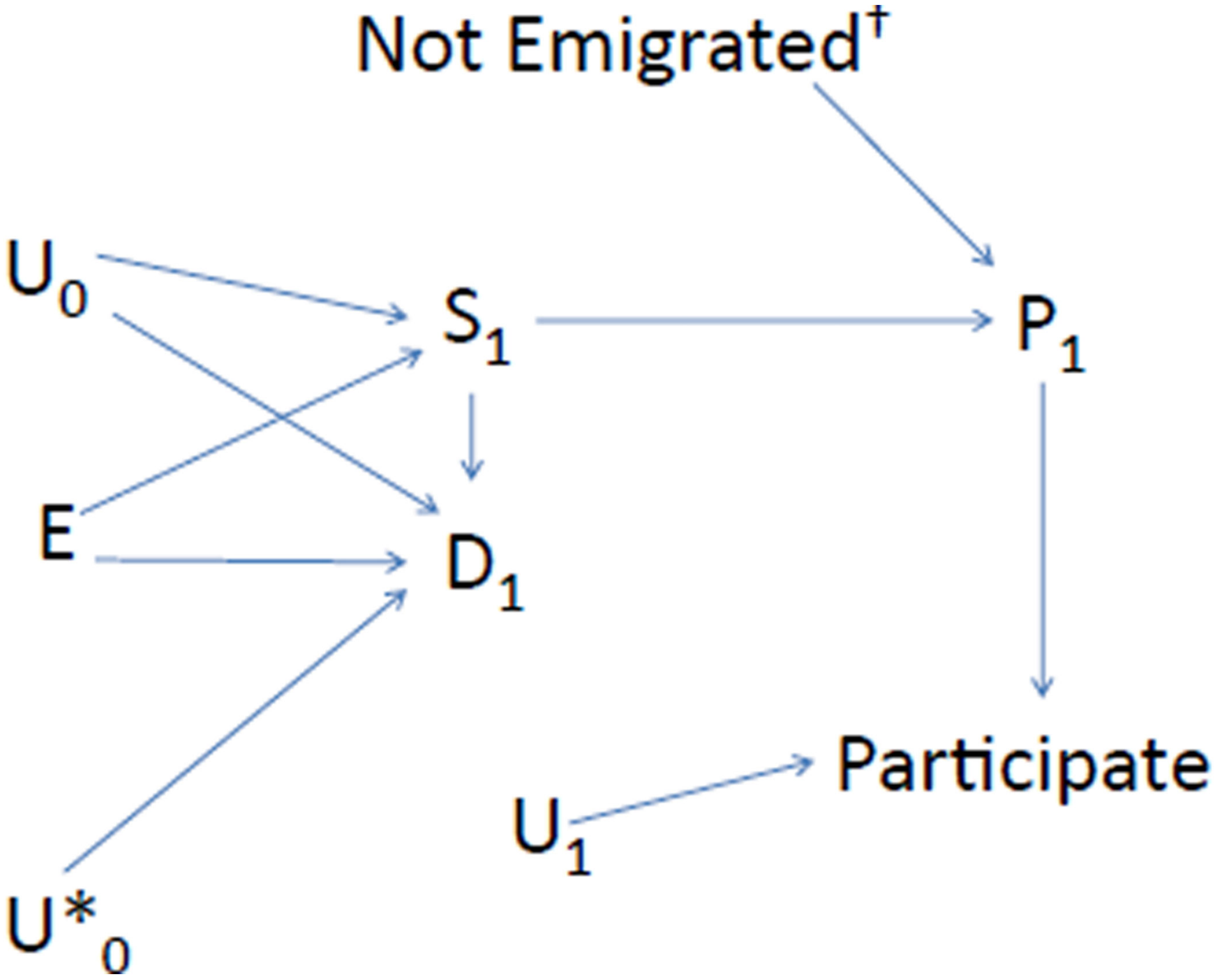

Fig. 1.

The figure summarizes causal relationships using a DAG for the baseline population P_0. If correct, estimator 2 should be unbiased (see text). D_1 represents disease presence at age a_1, U_0^* other causes of disease, U_0 other causes of survival (S_1), and disease (D_1). P_1 is the survey population; membership depends deterministically on survival, and not emigrating or other loss, but not directly on exposure E or D_1. Participation depends on P_1 and other factors U_1. †Emigration, other factors affect being in population P_1.

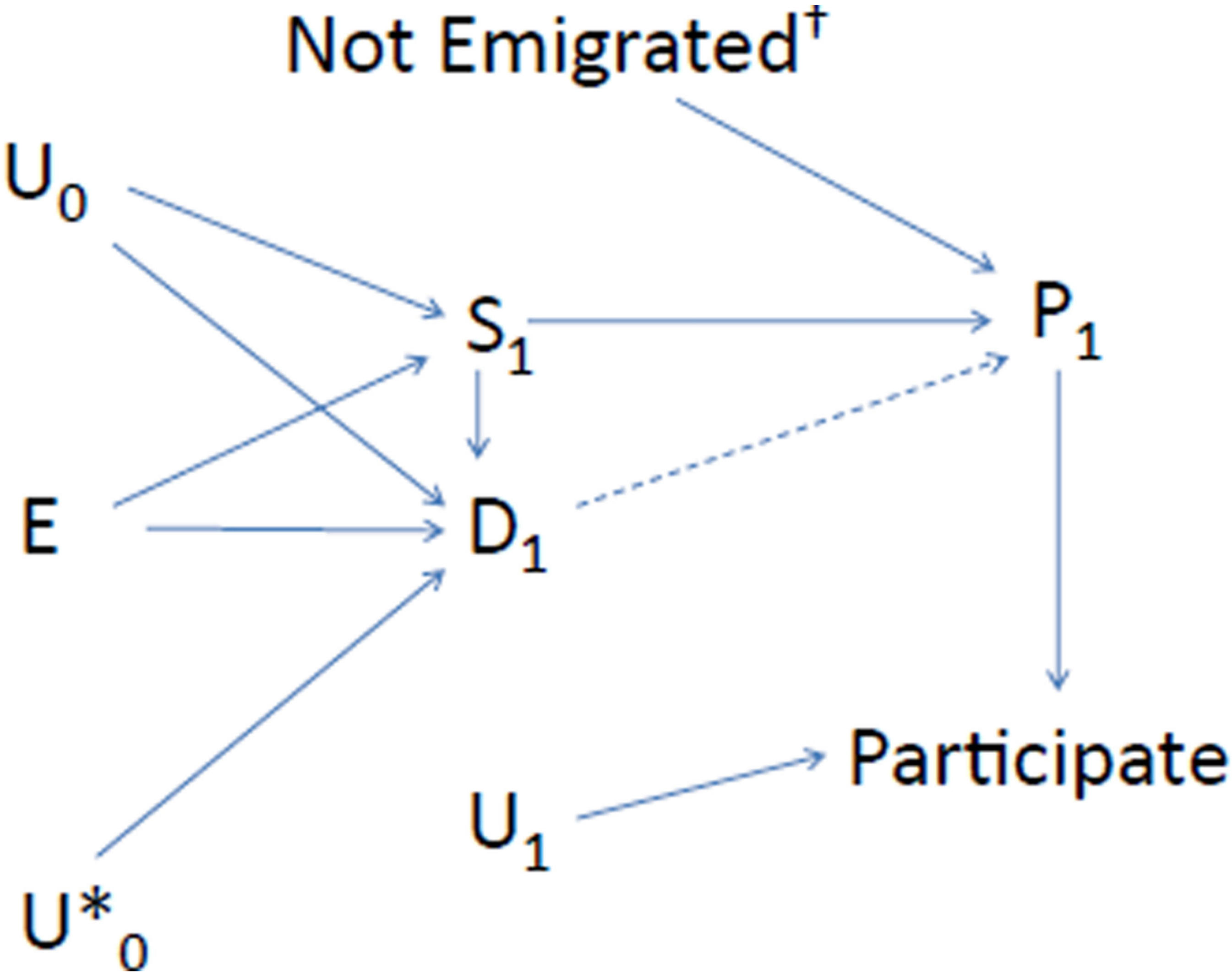

Fig. 2.

The figure illustrates a situation such as that in Figure 1, with an additional effect (dotted line) that could underlie bias in estimator 2. For example, bias is expected if S_1 and D_1 affect membership in P_1 (see text). D_1 represents disease presence at age a_1, U_0^* other causes of disease, U_0 other causes of survival (S_1), and disease. P_1 is the survey population; membership depends on survival, not emigrating, and D_1. Participation depends on P_1 and other factors U_1. †Emigration, other factors affect being in population P_1.

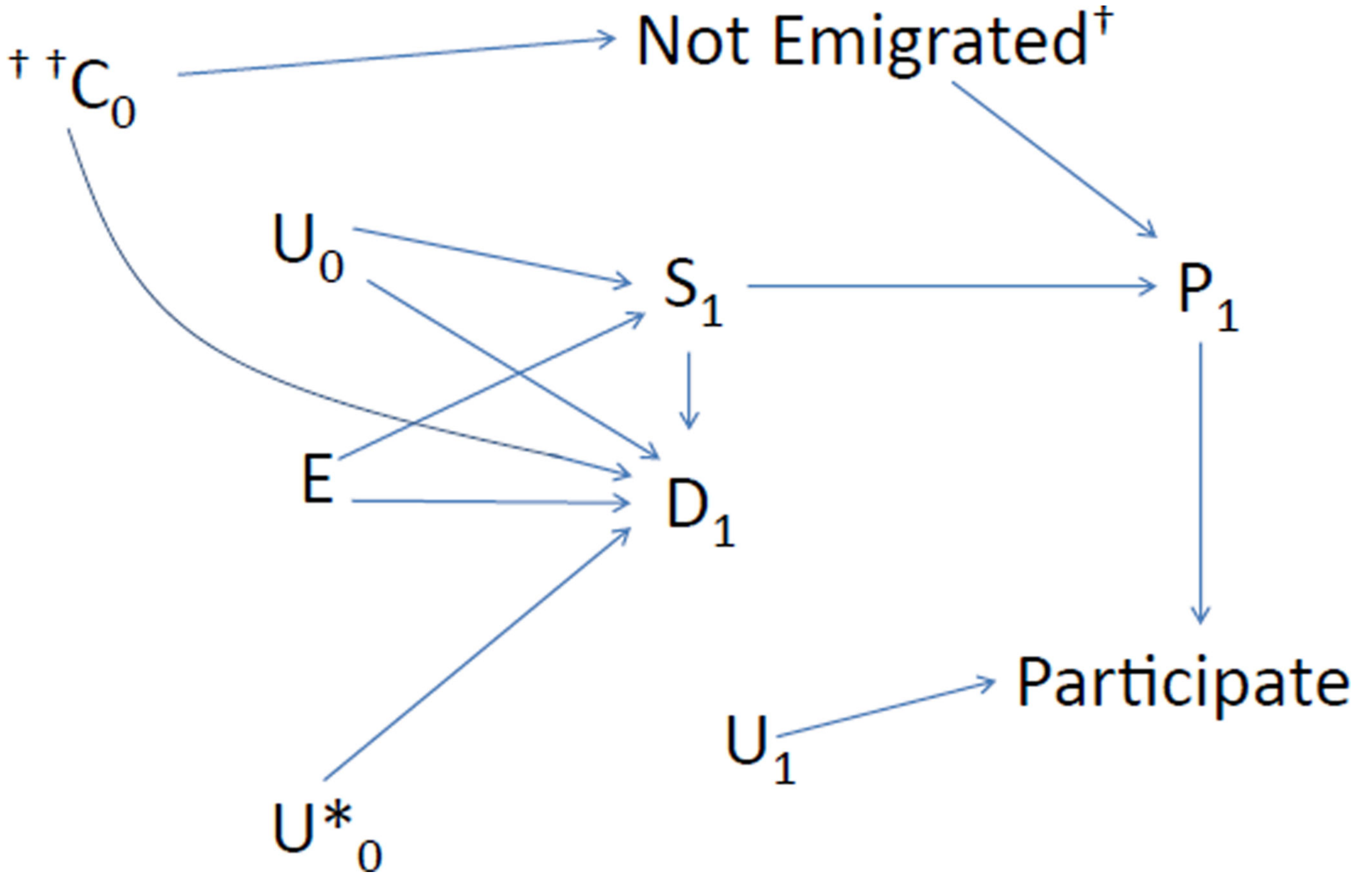

Fig. 3.

The figure illustrates a situation such as that in Figure 1, but with a common cause (C_0) of prevalent disease D_1 and emigration. Exposure-specific prevalences in the survey population (P_1) would be expected to differ from those in all survivors, and bias is expected. U_0^* represents other causes of disease, U_0 other causes of survival (S_1), and disease. P_1 is the survey population; membership depends deterministically on survival, not emigrating or other loss, and D_1. Participation depends on P_1 and other factors U_1. †Emigration, loss other factors affect being in population P_1. ††C_0 is common cause of prevalent disease and emigration, so prevalence in P_1 expected to differ from that in all survivors.

Table 1

Definitions and notation

Term	Brief definition
$[D_{i,a}(e), S_{i,a}(e)]$ counterfactual outcome vector	Components are counterfactual outcomes $D_{i,a}(e)$, $S_{i,a}(e)$ defined next. $[D_{i,a}(e), S_{i,a}(e)] = [1,1]$ if subject i is alive with disease at age a; = [0,1] if subject i is alive without disease at age; = [0,0] if subject i is not alive at age a; all if exposure had been set to e
$D_{i,a}(e)$—counterfactual disease outcome	First component of counterfactual-outcome vector $[D_{i,a}(e), S_{i,a}(e)]$, defined previously
$S_{i,a}(e)$—counterfactual survival, subject i	second component of counterfactual-outcome vector $[D_{i,a}(e), S_{i,a}(e)]$, defined previously
$D_{i,a}(1) - D_{i,a}(0)$	Individual causal effect on disease presence at age a
cPD—causal prevalence difference	Population-average effect of exposure at age a_0 on disease prevalence at a, in defined target population (Equation 1),
Exchangeability—disease	The counterfactual outcome with E set to e, is independent of actual exposure: $D_{i,a}(e) \amalg E_i$; exchangeability can be conditional on covariates C: $D_{i,a}(e) \amalg E_i \mid C$
Exchangeability—survival	The counterfactual outcome with E set to e, is independent of actual exposure: $S_{i,,a}(e) \amalg E_i$; exchangeability can be conditional on covariates C: $S_{i,a}(e) \amalg E_i \mid C$
Consistency	The observed outcome equals the counterfactual outcome if exposure were set to the actual exposure: $D_{i,a}(e)=D_{i,a}$ if $E_i=e$
Stable unit treatment value assumption	The outcome of individual i is independent of the exposure status of all other individuals: $D_{i,a}(e) \amalg E_j$ for $i \neq j$

Table 2

Examples of surveys for which the “parent” population P_0 may be specifiable[*]

Survey	Population P_1	Parent Population P_0
Population-based national survey (e.g. NHANES 3)	U.S. residents, noninstitutionalized, age a_1—excluding immigrants between age a_0 and a_1	U.S. residents, noninstitutionalized, age a_0, ($a_0 < a_1$)
Population-based statewide telephone ssurveys (e.g., BRFSS)	State residents, noninstitutionalized, age a_1—excluding immigrants between age a_0 and a_1	State residents, noninstitutionalized, age a_0, ($a_0 < a_1$)
Population-based national telephone survey (e.g., NHIS)	U.S. residents, noninstitutionalized, age a_1—excluding immigrants between age a_0 and a_1	U.S. residents, noninstitutionalized, age a_0, ($a_0 < a_1$)

BRFSS = Behavioral Risk Factor Surveillance System; NHANES = National Health and Nutrition Examination Survey; National Health Interview Survey (NHIS).

[*] P_0: as in the main text, P_1 is the population (age a_1) sampled for the survey, and P_0 is the parent population (age a_0) defined so that P_1 consists of all surviving members of P_0. To assure temporal precedence, a_0 is less than a_1.