medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Pathway-specific population attributable fractions

Maurice O'Connell[1] and John Ferguson [*1]

[1] *HRB Clinical Research Facility, National University of Ireland Galway*

[*] corresponding author

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Abstract

A population attributable fraction (PAF) represents the relative change in disease prevalence that one might expect if a particular exposure was absent from the population. Often, one might be interested in what percentage of this effect acts through particular pathways. For instance, the effect of excessive alcohol intake on stroke risk may be mediated by blood pressure, body mass index and several other intermediate risk factors. In this situation, attributable fractions for each mediating pathway of interest can be defined as the relative change in disease prevalence from disabling the effect of the exposure through that mediating pathway.

This quantity is related to, but distinct from the recently proposed metrics of direct and indirect PAF by Sjölander. In particular, while differing pathway-specific PAF will each usually be less than total PAF, they may sum over differing mediating pathways to more than total PAF, whereas direct and indirect PAF must sum to total PAF. Here, we present definitions, identifiability conditions and estimation approaches for pathway-specific attributable fractions. We illustrate results, and comparisons to indirect PAF using INTERSTROKE, a case-control study designed to quantify disease burden attributable to a number of known causal risk factors.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Introduction

In causal inference, in addition to estimating an overall causal effect of an exposure, one is often interested in the mediating pathways by which the exposure effects an outcome. For instance, the effect of excessive alcohol intake on stroke risk may be mediated by blood pressure, body mass index and several other intermediate risk factors; mediation analysis [1, 2, 3] tries to appropriate effect sizes to each of these known causal pathways.

In an analogous fashion, one can examine decompositions of attributable fractions into direct and indirect components. For example, one might ask 'what percentage of disease burden might be eliminated if the direct mechanisms by which an exposure effects disease were disabled?' Recently, Sjölander [4] proposed definitions and estimation approaches for population attributable fractions (PAF) in the spirit of the work by Robins, Greenland and Pearl that attempts to answer these questions. As we show later, Sjölander's definitions of direct and indirect PAF can be viewed as types of sequential PAF [5, 6], but formed by disabling causal pathways rather than eliminating risk factors. From this view, it turns out that direct and indirect PAF are defined in a somewhat asymetric manner, which may skew the relative disease burden appropriated to the differing pathways. Here, we propose an alternative metric, the pathway-specific PAF, that summarizes the disease burden associated with a particular causal pathway, which is more generalizable to multiple causal pathways and as opposed to Sjölander's approach treats all pathways in a symmetric fashion. In particular, while differing pathway-specific PAF will usually be less than total PAF, they may sum over differing

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

mediators to more than total PAF whereas direct and indirect PAF must sum to total PAF. In the next section, we present definitions, identifiability conditions and estimation approaches for pathway-specific attributable fractions. We illustrate results and comparisons to indirect PAF using data from INTERSTROKE, a case-control study designed to quantify disease burden attributable to a number of known causal risk factors for stroke. We finish the manuscript with a discussion, emphasizing differences between pathway-specific and indirect PAF and examining whether additivity to total PAF is a sensible property to require from our definitions.

Methods

Potential outcome notation used for mediation analyses

As is usual in the causal inference literature, we will define random variables for observed quantities with unscripted notation, whereas potential outcomes will be denoted using subscripts. In all cases, we use upper case letters to denote random quantities, and lower-case to denote quantities that are fixed or intervened on. In particular, let C denote a vector of confounders, $A \in \{0, 1\}$ a binary exposure of interest, M a mediator on the causal pathway from A to Y (as in Sjölander, [4], M can be binary, multi-category or continuous) and $Y \in \{0, 1\}$ a binary disease outcome. In the setting that there are several ($K > 1$) known mediators, we denote these M^1,...,M^K. Figure 1 below demonstrates a multi-mediator scenario with 3 mediators, M^1, M^2 and M^3.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

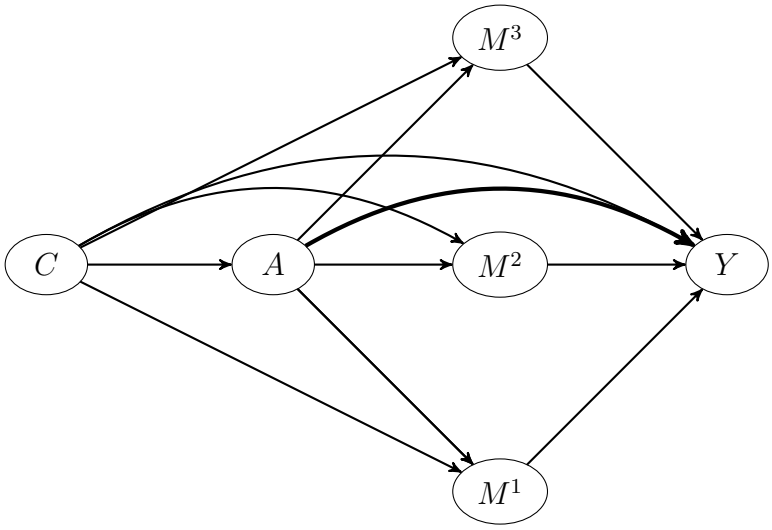

Figure 1: DAG showing causal structure linking exposure, mediators and outcome

Potential outcomes, assuming the exposure is set at level $a \in \{0, 1\}$ and the mediator is set at level m are denoted $Y_{a,m}$. In the case that there are $K > 1$ mediators, the corresponding potential outcome setting exposure to a and the mediators to m^1,...m^K is $Y_{a,m^1,...,m^K}$. One can also define counterfactuals for mediator k, assuming A is set to a as M_a^k, and nested counterfactuals for the outcome $Y_{a,M_{a*}}$, the last expression having the interpretation of the outcome that would be observed if A was set to a, and M was set to the value that it would naturally attain had A equalled $a*$. When $M_a \neq M_{a*}$ these nested counterfactuals are sometimes referred to as 'cross world' counterfactuals, since they imagine combinations of values of the exposure and mediator that could never occur together even under an intervention [7, 8]. We will sometimes abbreviate Y_{a,M_a} as Y_a.

As is usual with causal inference using the potential outcomes framework, we make Stable Unit Treated Value Assumptions (SUTVA) [9], which im-

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

plies that the relationship between potential and observed outcomes satisfy 'consistency', or that $Y = Y_{a,M_a}$ when $A = a$. In addition, mediation analysis requires some technical conditional independence assumptions which we will list as needed in the following sections.

Total PAF, direct PAF and indirect PAF

Despite being an intrinsically causal idea, attributable fractions were originally defined in a non-causal framework. Recently the following potential outcomes definition is becoming more prominent [10, 11, 12]:

$$PAF_{total} = \frac{P(Y = 1) - P(Y_0 = 1)}{P(Y = 1)}. \tag{1}$$

Note that if we think of Y_0 as the potential outcome for an individual if they were never exposed to the risk factor, (1) can be directly interpreted as the relative change in disease prevalence if an exposure was absent from the population. That is, for exposures such as tobacco usage, where the pattern of longitudinal exposure as well as current exposure is important in measuring disease risk, we interpret the attributable fraction as a comparison of current disease risk with counterfactual disease risk under the scenario that tobacco usage in the population was always zero.

Recently, Sjölander [4] introduced the ideas of mediation into the literature for PAF, defining a decomposition of the above total PAF into direct and indirect components, relevant for a single mediator. The direct PAF aims to measure the effect of the exposure directly on the response, that is the proportion of outcome events that can be attributed to direct and not medi-

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

ated pathways; likewise the indirect PAF aims to measure the proportion of outcome events that can be attributed to mediated effects. The definitions :

$$PAF_{direct} = \frac{P(Y=1) - P(Y_{0,M}=1)}{P(Y=1)} \tag{2}$$

$$PAF_{indirect} = \frac{P(Y_{0,M}=1) - P(Y_{0,M_0}=1)}{P(Y=1)}. \tag{3}$$

are constructed to enforce an additivity property (direct + indirect = total), paralleling the definitions of natural direct and indirect, given in [8] and [7]. From the definitions, it is clear that the direct PAF can be interpreted as what the PAF would be if the exposure did not effect the distribution of the mediator, although it is not immediately clear what interpretation the indirect PAF may have except being the remainder when the direct PAF is subtracted from the overall PAF.

Sjölander's Direct and Indirect PAF as types of sequential PAF

We could think of the effect of eliminating the risk factor A, as the effect of first disabling direct effects due to A (i.e. a comparison of $P(Y_{A,M}=1)$ and $P(Y_{0,M}=1)$) and afterwards disabling the remaining mediated effect of A, (i.e. a comparison of $P(Y_{0,M}=1)$ and $P(Y_{0,M_0}=1)$). The first operation corresponds to Sjölander's direct effect and the second to the indirect effect above. This interpretation suggests Sjölander's direct and indirect PAF are kinds of sequential PAF [5, 13], but defined on groups of pathways

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

rather than groups of risk factors and differing in their elimination position (here the direct pathway is removed first, corresponding to the direct PAF, the indirect pathway is subsequently removed, corresponding to the indirect PAF). Removal ordering can sometimes have substantial effects on the size of sequential PAF [6], particularly when independent mechanisms interact in causing disease. For instance, suppose for some individuals with $A = 1$, Y_{1,M_1}=1, but $Y_{0,M_1} = Y_{1,M_0} = Y_{0,M_0} = 0$). For these individuals, the contribution to the direct PAF $Y_{1,M_1} - Y_{0,M_1}$ from eliminating A is 1, and the contribution to the indirect PAF: $Y_{0,M_1} - Y_{0,M_0}$ is 0. This seems unsatisfactory as for these individuals the indirect and direct pathways interact equally in causing disease (in other language, the direct and mediating effects of A are a joint sufficient cause for $Y = 1$ [11]). In general, the indirect PAF defined above might be surprisingly small as it is an operator on the population where the direct effect of the exposure on the outcome has been disabled, and many disease cases that would occur in the regular population would have already been prevented in this hypothetical population. This motivates a new type of attributable fraction, specific to a particular causal pathway, which removes the asymetries in the definitions of direct and indirect PAF discussed above.

pathway-specific PAFs

The pathway-specific attributable fraction for the mediating pathway $A->M->Y$ in the example above is defined as

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

$$PAF_{A->M->Y} = \frac{P(Y=1) - P(Y_{A,M_0}=1)}{P(Y=1)} \tag{4}$$

This metric has several desirable properties. As above, this is a kind of sequential PAF on pathways, but now corresponding to disabling the effect of the mediating pathway first. The direct PAF defined by Sjölander, also being a sequential PAF but where the direct pathway is disabled first, is correspondingly a pathway-specific PAF:

$$PAF_{A->Y} = \frac{P(Y=1) - P(Y_{0,M}=1)}{P(Y=1)} \tag{5}$$

Whereas the PAF_{direct} and $PAF_{indirect}$ pertain to single mediated pathways, pathway-specific PAF can easily be extended to K>1 mediated pathways. For instance, if there are $K = 3$ known mediators, pathway-specific attributable fractions for each of the 4 pathways are defined as:

$$\begin{aligned} PAF_{A->M^1->Y} &= \frac{P(Y=1) - P(Y_{A,M_0^1,M^2,M^3}=1)}{P(Y=1)}, \\ PAF_{A->M^2->Y} &= \frac{P(Y=1) - P(Y_{A,M^1,M_0^2,M^3}=1)}{P(Y=1)}, \\ PAF_{A->M^3->Y} &= \frac{P(Y=1) - P(Y_{A,M^1,M^2,M_0^3}=1)}{P(Y=1)}, \\ PAF_{A->Y} &= \frac{P(Y=1) - P(Y_{0,M^1,M^2,M^3}=1)}{P(Y=1)}. \end{aligned} \tag{6}$$

The causal graph associated with $K = 3$ mediators is illustrated in Figure 1. While in this manuscript, we will continue to illustrate formulae and estimation approaches for the setting where $K = 3$, the methods extend in

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

a staightforward way for other values of K. Note that the last equation for $PAF_{A->Y}$, is the obvious extension of Sjölander's direct effect to multiple mediators. While this formula seems to differ slightly from the other pathway-specific PAFs, this difference is slightly artificial. For instance, if we regard the direct pathway $A->Y$ as operating completely through a set of unobserved mediators M^4 (extending the potential outcome notation, we would require that $Y_{0,m^1,m^2,m^3,m^4} = Y_{1,m^1,m^2,m^3,m^4}$, for all possible fixed values m^1, m^2, m^3, m^4), we could then write $PAF_{A->Y}$=$PAF_{A->M^4->Y}$, with the formula for $PAF_{A->M^4->Y}$ then appearing symmetrical with the formulae for the other mediating pathways.

Identifiability conditions

Further identifiability conditions are needed to estimate (6) when values on the exposure, mediators, observed confounders and outcome are available. In particular, we need to assume:

1. $Y_{a,m^1,m^2,m^3} \perp\!\!\!\perp (M^1_{a_1}, M^2_{a_2}, M^3_{a_3})|A, C$ for any choice of $(a_1, a_2, a_3) \in \{0,1\}^3$ and possible values (m^1, m^2, m^3) for the mediator

2. $A \perp\!\!\!\perp M_a|C$, for $a \in 0, 1$

3. $(M^1_{a_1} \perp\!\!\!\perp M^2_{a_2} \perp\!\!\!\perp M^3_{a_3}|A, C)$ for any choice of $(a_1, a_2, a_3) \in \{0,1\}^3$

Under these assumptions, we show in the supplementary material that:

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

$$PAF_{A->M^1->Y} = (P(Y=1) - E_{P^1}(P(Y=1|A,C,M^1,M^2,M^3)))/P(Y=1),$$
$$PAF_{A->M^2->Y} = (P(Y=1) - E_{P^2}(P(Y=1|A,C,M^1,M^2,M^3)))/P(Y=1),$$
$$PAF_{A->M^3->Y} = (P(Y=1) - E_{P^3}(P(Y=1|A,C,M^1,M^2,M^3)))//P(Y=1) \tag{7}$$

where E_{P^1} is the expectation operation with respect to a modified distribution of (A, C, M^1, M^2, M^3), that is constructed via sampling (A, C, M^2, M^3) from its marginal observational distribution, and then sampling M^1 from the conditional observational distribution of M^1 given A=0 and C. P^2 and P^3 are constructed similarly.

Similarly for the pathway-specific direct effect:

$$PAF_{A->Y} = (P(Y=1) - E(P(Y=1|A=0,C,M^1,M^2,M^3)))/P(Y=1) \tag{8}$$

where E is now the regular expectation operator with respect to the observational distribution of (A, C, M^1, M^2, M^3).

Estimation

In cohort and cross-sectional studies, the differing components of (7) and (8), as well as the distribution being integrated over can be estimated without bias from the empirical conditional expectations and distributions in the original data. For case control studies, we use a simple-reweighting trick which assumes that the prevalence of disease, π is known, and the sampled disease

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

cases and controls are randomly selected from their respective populations. We assume for simplicity that the case:control matching ratio is 1:r, for some $r \geq 1$. Under these assumptions, the components of (7) and (8) can be found as the corresponding empirical expectations and distributions in the re-weighted dataset where cases are assigned weights $w_i = 1$, and controls are assigned weights $w_i = (1/\pi - 1)/r$. (Note that under the assumptions that prevalence is known and the cases and controls are randomly selected from their source populations, this reweighted sample is a random sample from the source population). Suppose then, that the researcher specifys and estimates correct models for $P(Y = 1|A, C, M^1, M^2, M^3)$ and a correct conditional model for each M^k, conditioned on A and C (perhaps in the re-weighted population if the original data is from a case-control study). To avoid numerical integration in the continuous mediator case, and simplify the algorithm, we propose that $E_{P^1}(P(Y = 1|A, C, M^1, M^2, M^3))$ is estimated via the following algorithm:

1. Choose a number of simulation iterates, S
2. for(j in 1:S){
   (a) For each individual in the data, i, with exposure, A_i, covariate vector C_i and mediators M^{1i}, M^{2i}, M^{3i} simulate M^{1*} from the estimated conditional distribution of M^1 given A_i and C_i
   (b) Estimate $P(Y = 1|A_i, C_i, M^{1*}, M^{2i}, M^{3i})$ for each i using the estimated statistical model
   (c) Calculate $E_j = \sum_{i \leq N} P(Y = 1|A_i, \widehat{C_i, M^{1*}}, M^2, M^3)/N$ for cohort/cross sectional designs; calculate

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

$E_j = \sum_{i \leq N} w_i P(Y = 1 | A_i, C_i, \widehat{M^{1*}, M^{2i}}, M^{3i}) / \sum_{i \leq N} w_i$ for case control designs.

}

3. $E_{P^1}(P(Y = 1 | \widehat{A, C}, M^1, M^2, M^3)) = \sum_{j \leq S} E_j / S$

$E(P(Y = 1 | A = 0, C, M^1, M^2, M^3)))$ can be estimated more simply by estimating $E(P(Y = 1 | A = 0, \hat{C}_i, M^{1i}, M^{2i}, M^{3i}))$ for each individual, and averaging over individuals in the data, taking care to incorporate weighting under a case control design. As an alternative, a double robust estimator for $E(P(Y = 1 | A = 0, C, M^1, M^2, M^3))$ can be derived using the same approaches Sjölander describes in [4].

Data Example

INTERSTROKE, [14], is a large international case control study designed to quantify the contribution of established risk factors to stroke prevalence at a global level. Here we consider investigate the possible mediating effects of physical activity on stroke through waist hip ratio, apolipoproteins B to A1 ratio and diagnosis of high blood pressure. We treat waist hip ratio and apolioproteins as continuous variables, whereas diagnosed high blood pressure is binary. Confounders and mediators are as shown in the causal structure shown by Figure 2. To estimate Sjölander's direct and indirect attributable fractions, and the pathway-specific attributable fractions described above, we fit a main-effects logistic regression predicting stroke status as a function of age,sex,diet,physo-social stress factors,smoking status, alcohol use, physical

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

activity, waist hip ratio, apoB/apoA ratio and clinically diagnosed high blood pressure, with the terms for wasit hip ratio and apoB/apoA ratio entering as 5-degree of freedom natural cubic splines. In this regression, stroke controls were upweighted by a factor of 284 to reflect a yearly stroke incidence of first stroke of 0.0035 or 3.5 strokes per 1000 individuals per year, estimated via data from the global burden of disease [15]. To apply the algorithm in the previous section, we also need to simulate from the predicted distribution for the mediators, ApoB/ApoA ratio, waist hip ratio and clincially diagnosed hypertension (conditioned on values for age, sex, region, physical activity, diet-score, stress, smoking and alcohol). To do this, we resample residuals from a fitted linear model for ApoB/ApoA ratio and waist hip ratio and add to resampled residuals to the fitted values, whereas for hypertension we simply draw Bernoulli variables with probabilities according to the fitted logistic models.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Pathway	total PAF	Direct (overall)	Direct (Sjölander)	Indirect (Sjölander)	Path specific
PHYS->HBP->STROKE	0.40 (0.34, 0.47)	0.35 (0.27, 0.42)	0.39 (0.33, 0.46)	0.013 (-0.014, 0.039)	0.041 (0.014, 0.068)
PHYS->WHR->STROKE	0.40 (0.34, 0.47)	0.35 (0.27, 0.42)	0.40 (0.33, 0.47)	0.005 (-0.016, 0.026)	0.032 (0.007, 0.058)
PHYS->APOB->STROKE	0.40 (0.34, 0.47)	0.35 (0.274, 0.42)	0.39 (0.32, 0.45)	0.018 (-0.005, 0.041)	0.045 (0.016, 0.073)

Table 1: Results from estimating the pathway-specific mediated and direct effects on the INTERSTROKE dataset using $S = 100$, with standard errors calculated using 200 bootstrap iterations. The 95% confidence interval is in brackets (mean estimate $\pm 1.96 *$standard deviation). PAF estimates are rounded to 2 significant digits

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

Sjölander's direct PAF is defined on a mediator by mediator basis in the following fashion:

$$PAF_{direct,M^1} = \frac{P(Y=1) - P(Y_{0,M^1,M_0^2,M_0^3} = 1)}{P(Y=1)} \tag{9}$$

and is estimated in a similar fashion to the methods for the pathway-specific PAF. Note that while there is only one overall direct pathway-specific PAF, Sjölander's definition is with reference to a single mediating pathway - and will change depending on that pathway. For instance, in this example the Sjölander direct effect with reference to high blood pressure essentially pools together all other pathways not going through high-blood pressure. In this example, the mediating pathways and direct pathway due to lack of exercise are all deleterious, and one would expect the pathway-specific attributable fraction to be somewhat larger than Sjölander's indirect PAF, which is what we see in Table 1 - the estimated pathway-specific PAFs ranging from 3.2%-4.5%, whereas the indirect PAFs range from 0.5% to 1.8%. On the other hand, the overall direct-pathway-specific PAF is 35%. The analysis suggests that population disease burden for stroke attributable to physical activity may partially depend on these mediating pathways, but mostly depends on other mechanisms not directly through blood pressure, waist hip ratio or lipids. As with any causal analysis, these tentative conclusions depend jointly on correct modeling of conditional probability distributions and on the validity of the causal identifiability assumptions that we listed earlier. These assumptions are doutful here: the modeling here which thresholds blood pressure and ignores potential exposure-mediator interactions is po-

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

tentially overly-simplistic, and it is probable that unmeasured variables exist (i.e. variables not listed in Figure 2) which confound the exposure mediator or mediator outcome relationship. However, the differences between the estimated pathway-specific direct and mediated PAFs observed here are so large that it would be hard to completely attribute these differences to incorrect modeling or an incomplete causal model.

Discussion

In this paper we have introduced pathway-specific attributable fractions as a metric for measuring the disease burden attributable to particular exposure mediator pathways. As shown in Table 2, this idea is related to but distinct from the recently proposed metrics of direct and indirect PAF [4] in a number of ways. All the metrics can be interpreted as kinds of sequential PAF defined on pathways. In particular, one can informally interpret a pathway-specific PAF for a mediated pathway as the relative change in disease burden if a particular mediated pathway were disabled. In contrast one could think of the indirect PAF as associated with disabling the mediating pathway, but this time subsequent to disabling the direct pathway. Since the effect of disabling both the direct and mediating pathways is equivalent to the effect of eliminating the risk factor, this effectively forces the additivity property that total PAF is the sum of direct and indirect PAF.

While this additivity property at first seems appealing, it perhaps is unnatural in the context of attributable fractions, where it is well recognized that the PAF for differing risk factors may sum to more than the joint PAF

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

and sometimes to more than 1 [16]. Indeed, the lack of additivity for attributable fractions can be understood again from the context of sequential PAFs. The usual 'non-sequential' definition of PAF for a risk factor can be interpreted as the sequential PAF for that risk factor if it is eliminated first in sequence before other risk factors. While a set of sequential PAFs, corresponding to eliminating a group of risk factors in a certain order must sum to the joint PAF for that set of risk factors, these sequential PAFs generally decrease as a function of elimination position with the result that non-sequential PAFs generally sum to more than the joint PAF. The sufficient/component cause framework [17] gives a simple but enlightening explantion for this phenomenon. For particular individuals, a certain collection of risk factors (perhaps diet, stress and tobacco usage) might collectively lead to disease at a particular point in time, but the disease may not have occured at that time if any of the risk factors were not present. The same logic implies that pathway-specific PAFs will tend to be larger than indirect PAFs as illustrated in this manuscript. We note that a larger result is not in itself a compelling reason to introduce a new metric; what constitutes a good definition of a causal effect or estimand is open to some interpretation. However, compared to indirect PAF, pathway-specific PAFs are defined more in the spirit of regular attributable fractions, and in a symmetric manner for direct and mediated pathways. As a result, an examination of pathway-specific PAFs will constitute a fairer reflection of the relative strengths of direct and mediating pathways compared to a similar comparison of direct and indirect PAF.

Note that the issue of whether enforcing additivity for PAFs is sensible

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

is paralleled to a certain extent in the wider literature regarding mediation. Natural direct and indirect causal effects have been defined by Pearl [18] which don't necessarily add up to the total causal effect, and redefined by authors such as Vanderweele [8] so that they do. It turns out that for linear systems with no interactions (where the conditional expectation of the mediator given confounders and exposure is linear in confounders and exposure, and similarly the conditional expectation of the outcome is linear in confounders, mediator and exposure), all definitions will co-incide. In a similar way pathway-specific PAF and indirect PAF will agree in this very specific setting, although for a binary disease outcome linearity or even approximate linearity in the conditional expectation of the outcome would be implausible, particular if the mediator is continuous.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

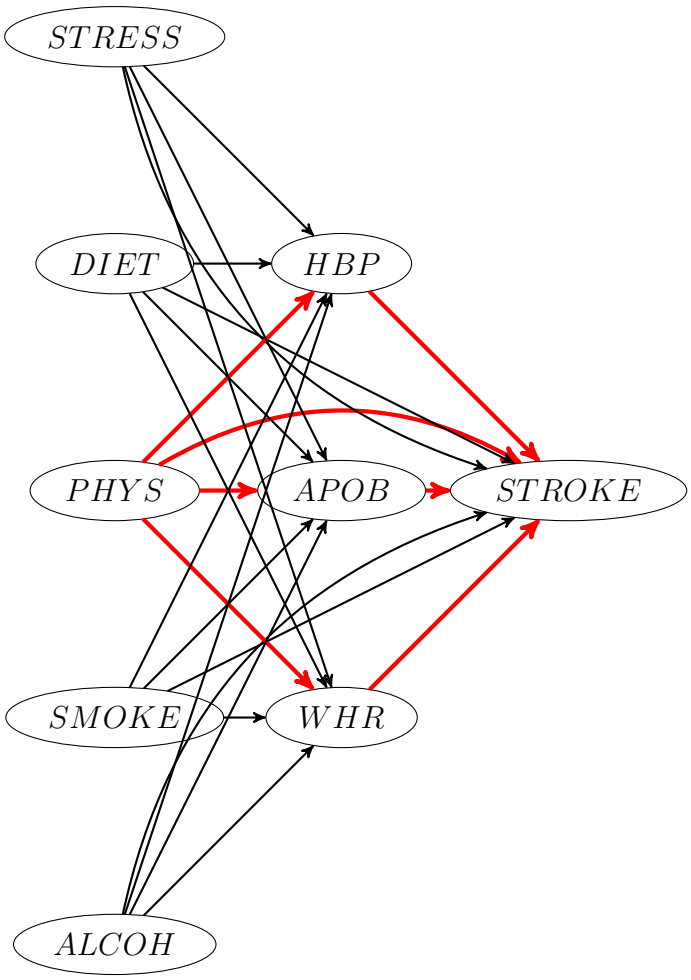

Figure 2: DAG showing assumed causal structure for risk factors in INTERSTROKE. The direct and mediating pathways associated with physical activity are highlighted in bold red. Age, sex and geographic region are confounders for the risk/factor disease relationship for all listed factors on the figure.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

	Indirect PAF	**pathway-specific PAF**
Number of mediators	Focused on one mediator, as the difference between total PAF and direct PAF	Naturally Focused on several mediators
Interpretation of direct PAF	Defined in context of a single mediator as what the PAF would be if the exposure didn't affect the distribution of the mediator	Represents the PAF attributable to none of the measured mediators. Can be regarded as a pathway-specific PAF for all unobserved mediating pathways
Interpretation of mediated PAF	The extra attributable risk that is left over from the total PAF after subtracting the direct PAF	The disease burden associated with a hypothetical intervention leaving the risk factor at its natural value but setting the mediator at the value that would be attained if the risk factor was eliminated)
Additivity	Indirect PAF+direct PAF =overall PAF by definition	Additivity not present by default.
Relative size of mediated PAF	Might be surprisingly small due to interpretation as seqential PAF corresponding to disabling the indirect pathway sequentially after the direct pathway already disabled	Usually larger than indirect PAF

Table 2: Comparison of pathway-specific attributable fractions with indirect and direct PAF

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

References

[1] R. M. Baron and D. A. Kenny, "The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations.," *Journal of personality and social psychology*, vol. 51, no. 6, p. 1173, 1986.

[2] J. M. Robins and S. Greenland, "Identifiability and exchangeability for direct and indirect effects," *Epidemiology*, pp. 143–155, 1992.

[3] J. Pearl, "Direct and indirect effects," *arXiv preprint arXiv:1301.2300*, 2001.

[4] A. Sjölander, "Mediation analysis with attributable fractions," *Epidemiologic Methods*, vol. 7, no. 1, 2018.

[5] G. E. Eide and O. Gefeller, "Sequential and average attributable fractions as aids in the selection of preventive strategies," *Journal of clinical epidemiology*, vol. 48, no. 5, pp. 645–655, 1995.

[6] J. Ferguson, M. O'Connell, and M. O'Donnell, "Revisiting sequential attributable fractions," *Archives of Public Health*, vol. 78, p. 67, July 2020.

[7] J. Pearl, M. Glymour, and N. P. Jewell, *Causal inference in statistics: A primer*. John Wiley & Sons, 2016.

[8] T. VanderWeele, *Explanation in causal inference: methods for mediation and interaction*. Oxford University Press, 2015.

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

[9] D. B. Rubin, "Randomization analysis of experimental data: The fisher randomization test comment," *Journal of the American Statistical Association*, vol. 75, no. 371, pp. 591–593, 1980.

[10] E. Suzuki, E. Yamamoto, and T. Tsuda, "On the relations between excess fraction, attributable fraction, and etiologic fraction," *American journal of epidemiology*, vol. 175, no. 6, pp. 567–575, 2012.

[11] M. A. Hernán and J. M. Robins, "Causal inference: what if," *Boca Raton: Chapman & Hill/CRC*, vol. 2020, 2020.

[12] J. Ferguson, N. O'Leary, F. Maturo, S. Yusuf, and M. O'Donnell, "Graphical comparisons of relative disease burden across multiple risk factors," *BMC medical research methodology*, vol. 19, no. 1, p. 186, 2019.

[13] J. Ferguson, A. Alvarez-Iglesias, J. Newell, J. Hinde, and M. O'Donnell, "Estimating average attributable fractions with confidence intervals for cohort and case–control studies," *Statistical methods in medical research*, vol. 27, no. 4, pp. 1141–1152, 2018.

[14] M. J. O'Donnell, S. L. Chin, S. Rangarajan, D. Xavier, L. Liu, H. Zhang, P. Rao-Melacini, X. Zhang, P. Pais, S. Agapay, *et al.*, "Global and regional effects of potentially modifiable risk factors associated with acute stroke in 32 countries (interstroke): a case-control study," *The lancet*, vol. 388, no. 10046, pp. 761–775, 2016.

[15] "Gbd compare data visualization, seattle, wa: Ihme, university of washington, 2018. available from http://vizhub.healthdata.org/gbd-compare,"

medRxiv preprint doi: https://doi.org/10.1101/2020.10.15.20212845; this version posted October 20, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.
All rights reserved. No reuse allowed without permission.

[16] A. K. Rowe, K. E. Powell, and W. D. Flanders, "Why population attributable fractions can sum to more than one," *American journal of preventive medicine*, vol. 26, no. 3, pp. 243–249, 2004.

[17] K. J. Rothman, "Causes," *American journal of epidemiology*, vol. 104, no. 6, pp. 587–592, 1976.

[18] J. Pearl, *Chapter in Book: C. Berzuini, P. Dawid, and L. Bernardinelli (Eds.), Causality: Statistical Perspectives and Applications*, ch. 12. The Mediation Formula: A guide to the assessment of causal pathways in nonlinear models, pp. pp. 151–179. 2012.