Causal attribution fractions for epidemiological studies, applied to a UK Biobank study of smoking and BMI

Anthony J. Webster

Nuffield Department of Population Health, Big Data Institute, Old Road Campus, University of Oxford, Oxford, OX3 7LF, UK.

Epidemiological studies often use proportional hazard models to estimate associations between potential risk factors and disease risk. It is emphasised that when the "backdoor criteria" from causal-inference applies, if diseases are sufficiently rare, then the proportional hazard model can be used to estimate causal associations. When the "frontdoor criteria" applies (allowing causal estimates with unmeasured confounders), similar estimates are found to mediation analyses with measured confounders. Reasons for this are discussed. An attribution fraction is constructed using the average causal effects (ACE) of exposures on the population, and simple methods for its evaluation are suggested. It differs from the attribution fraction used by the World Health Organisation (WHO), except for specific circumstances where the latter can agree or provide a bound. A counterfactual argument determines an individual's attribution fraction A_f in terms of proportional hazard estimates, as $A_f = 1 - 1/R$, where R is an individual's relative risk. Causally meaningful attribution fractions cannot be constructed for all known risk factors or confounders, but there are important cases where they can. As an example, systematic proportional hazards studies with UK Biobank data estimate the attribution fractions of smoking and BMI for 226 diseases. The attribution of risk is characterised in terms of disease chapters from the International Classification of Diseases (ICD-10), and the diseases most strongly attributed to smoking and BMI are identified. The result is a quantitative characterisation of the causal influence of smoking and BMI on the landscape of disease incidence in the UK Biobank population.

Introduction

The original aim of this research was to quantify the modifiability of disease risks by several behaviours and physical characteristics, such as smoking and body mass index (BMI). Risks are often quantified with attributable fractions and relative risks [1, 2], that are estimated using proportional hazard models [2, 3]. However, there are several ways of defining and estimating attributable fractions [2, 4, 5], and estimates of relative risks do not generally have a causal interpretation [2].

The recently established theory of causal inference [6, 7] is used to identify common situations where relative risk estimates from proportional hazard models provide estimates of causal associations. These can then be used to form attributable fractions that characterise the causal influence of exposures on disease risk. A causal analysis limits the exposures whose associations can be given a causal interpretation. Fortunately these included smoking and BMI, that are widely acknowledged to have a strong influence on a wide range of diseases.

Whereas statistics is the science of finding and describing patterns in data, epidemiology is the science of using statistics to make correct inferences. Although epidemiologists are careful to describe their results in terms of "associations", the purpose of epidemiology is to detect and quantify causal associations, e.g. between lifestyles and health [2, 8]. Recently the science of causal inference [6, 7, 9], has developed to allow estimates of causal associations to be made, without data from randomised control trials (RCTs). By exploring the relationship between causal estimates that are made with the "backdoor criteria" and the "do" calculus [6, 7], and conventional epidemiological estimates of relative risks using proportional hazards [3], it is shown that if formulated and interpreted correctly, many conventional epidemiological studies [2, 3] will correctly estimate causal associations. It is also shown how proportional hazards estimates can be applied to situations that satisfy the "frontdoor" criteria, and how the results can coincide with mediation analyses [9]. An attribution fraction is suggested that uses estimates of causal associations. It is closely related to the average causal effect (ACE) [6, 7], and it can agree with existing expressions when these are combined with estimates of causal associations [1, 4]. An attribution fraction for an individual is also formulated using a counterfactual argument that combines the "effect of treatment on the treated" [6, 7] and proportional hazard estimates, that gives a simple and widely used expression in terms of an individual's relative risk.

As an example of the approach with importance for public health policy, the attribution of disease incidence to smoking and BMI was considered for the UK Biobank cohort of over 500,000 UK men and women [10]. Attribution fractions for smoking and BMI were estimated for 226 diseases with statistically significant associations after adjusting for multiple

testing, using conventional proportional hazards models that are adjusted for known confounding factors. The results emphasise the heterogeneous influence of risk factors, with protective associations for several diseases. They also emphasise the pathogenic influence of smoking and BMI on some diseases, with 11 having an attributable fraction of 0.5 or greater.

Causal inferences

Relative risks and the "backdoor criteria"

Consider how disease risk may be influenced by the established common risk factors of education, socio-economic status, BMI, alcohol, and smoking (figure 1). It seems likely that for many diseases, education and socio-economic status could influence disease risk through the modifiable risk factors of BMI, alcohol, and smoking, in addition to any direct risk. In those circumstances education and socio-economic status are confounders (denoted with a vector Z), that influence both disease risk and the values of BMI, alcohol, and smoking (denoted with a vector X). For this causal model illustrated in figure 1, it is possible to estimate the consequences of setting BMI, alcohol, and smoking to a specific value $X = x$, corresponding to do$(X = x)$ using the "do" notation of Pearl [6, 7]. The situation corresponds to the well-known situation described by the

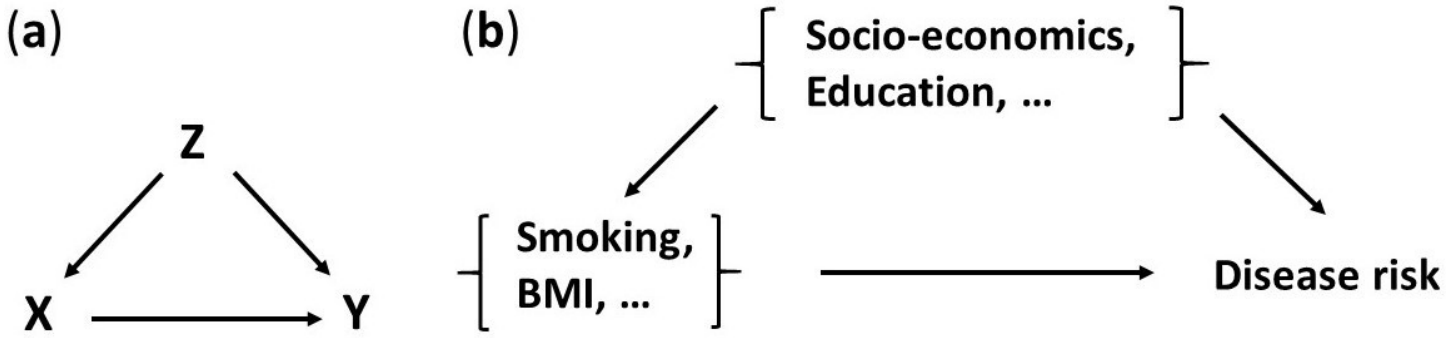


Figure 1: Consider the influence of one or more exposures X, on diseases Y, with confounding variables Z that satisfy the "backdoor criteria" [6, 7] (figure a). For example, X might include BMI, alcohol, and smoking, with confounders Z of socio-economic status and education (figure b).

"adjustment" formula, that states,

$$P(Y = y|\text{do}(X = x)) = \sum_z P(Y = y|X = x, Z = z)P(Z = z) \tag{1}$$

where for continuous variables, the sums are treated as integrals. The formula is constructed to account for the confounding influence of Z on both X and disease risk, and differs from that given by conventional probability theory, that would have $P(Z = z|X = x)$ instead of $P(Z = z)$. Take $Y = 1$ to denote presence of disease, and $Y = 0$ its absence, so that,

$$P(Y = 1, T < t|X = x, Z = z) = F(t, x, z) \tag{2}$$

where $F(t, x, z)$ is the distribution function (with covariates x, z), so that,

$$\begin{aligned} F(t,x,z) &= 1 - S(t,x,z) \\ &= 1 - \exp(-H(t,x,z)) \\ &\simeq H(t,x,z) \\ &= e^{\eta_x + \eta_z} H_0(t) \quad \text{with} \quad H_0(t) = \textstyle\int_0^t h_0(s)ds \end{aligned} \tag{3}$$

where in going from the 2nd to 3rd line we assume sufficiently rare diseases, as is the case for the first hospital admission for most diseases in UK Biobank [10, 12], and in going from the 3rd to the 4th lines we assume that the proportional hazards assumption [3] is valid for the disease being studied, with η_x and η_z being the linear predictor functions[1] for the (possibly)

[1]For a particular $X = x$, the linear predictor function is sometimes referred to as the "linear component", "risk score", or "prognostic index" [3].

multivariate variables x and z. For a probability density $f = dF/dt$ and hazard function $h = f/S$, $H(t) = \int_0^t h(t)$ is the cumulative hazard, and a proportional hazards model assumes that $h(t, x, z) = h_0(t)e^{\eta_x + \eta_z}$. Now using Eq. 1,

$$\begin{aligned} &P(Y = 1, T < t|\text{do}(X = x)) \\ &= \textstyle\sum_z P(Y = y, T < t|X = X, Z = z)P(Z = z) \\ &\simeq \textstyle\sum_z e^{\eta_x + \eta_z} H_0(t)P(Z = z) \\ &= e^{\eta_x} H_0(t)A_Z \end{aligned} \tag{4}$$

with $A_Z \equiv \sum_z e^{\eta_z} P(Z = z)$. This allows the incidence rates to be calculated for a (possibly hypothetical) situation where we have intervened in some way to set $X = x$, in terms of a baseline hazard function that is estimated in the usual way, using observational data in which Z can be correlated with both X and disease risk. Note that $P(Z = z)$ and $P(X = x)$ are implicitly the population values at the study's start.

At the baseline values of $x = x_0$ and $z = z_0$, by definition $\eta_x(x_0) = 0$ and $\eta_z(z_0) = 0$, so Eq. 4 gives $P(Y = 1, T < t|X = x_0, Z = z_0) = H_0(t)$. Then with the same approximations used to derive Eq. 4, Eq. 4 can be written in several different ways, for example with,

$$\begin{aligned} &P(Y = 1, T < t|\text{do}(X = x)) \\ &= e^{\eta_x} A_Z H_0(t) \\ &= e^{\eta_x} A_Z P(Y = 1, T < t|X = x_0, Z = z_0) \\ &= A_Z P(Y = 1, T < t|X, Z = z_0) \end{aligned} \tag{5}$$

When education and socio-economic factors are represented by Z, then the factor A_Z accounts for changes in risk due to both socio-economic factors and education, and the influence of setting $X = x$ is calculated through the factor e^{η_x}. If we could set X equal to the baseline values x_0, the probability distribution would be proportional to the baseline hazard function $H_0(t)$, amplified or shrunk by the factor A_Z. If the baseline values corresponded to the lowest disease risk, then $A_Z H_0(t)$ would be the lowest possible disease incidence rate that could have been achieved through lifestyle changes. Eq. 5 can be written as,

$$\frac{P(Y = 1, T < t|\text{do}(X = x))}{P(Y = 1, T < t|\text{do}(X = x_0))} = e^{\eta_x} \tag{6}$$

This gives a relative risk of disease within time t for a population with $X = x$, compared with a population with baseline values of $X = x_0$, in terms of estimates from a relative risk from observational studies, that have,

$$e^{\eta_x} = \frac{h(t|X = x, Z = z_0)}{h(t|X = x_0, Z = z_0)} \tag{7}$$

Appendix A relates the argument to probability densities and hazard functions.

The analysis above applies more generally to other examples, and to studies other than those involving disease or health. Similar results will apply whenever $F(t, x, z)$ can be factored as $H_0(t)g(x)q(z)$, for some functions $g(x)$ and $q(z)$, as was possible here because we consider a proportional hazards model and situations whose the incidence is sufficiently rare that we can approximate $F(t, x, z) \simeq H(t, x, z)$.

Unmeasured confounders and mediation - the "frontdoor criteria"

Another important result from causal inference, is the "frontdoor criteria" [6, 7]. A well-known example [7] is assessing the influence of smoking on disease risk in the presence of *unmeasured* confounders that influence both smoking use and disease risk, by using an additional measurement of tar in peoples' lungs (figure 2). Again we consider the adjustment formula for this situation in the limit of rare diseases, as above, and consider the simple specific example with continuous variables for e.g. average number of cigarettes per day and tar content of lungs. Although the estimated incidence rates will differ from those using proportional hazards models, the causal estimate for the influence of smoking on lung cancer, is the same as we might (with hindsight) have anticipated from mediation studies.

For the situation described in figure 2, the "front door" adjustment formula states [6, 7],

$$P(Y = y|\text{do}(X = x)) = \sum_z \sum_{x'} P(Y = y|Z = z, X = x')P(X = x')P(Z = z|X = x) \tag{8}$$

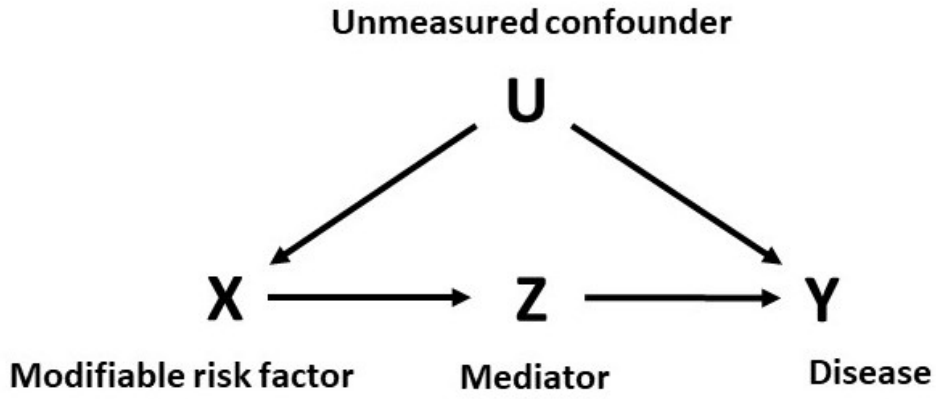


Figure 2: The "frontdoor criteria" estimates the causal influence of an exposure $\mathrm{do}(X = x)$, that is mediated by Z, in the presence of unmeasured confounders U that influence both the disease risk and the exposure X.

Using this, and proceeding as before,

$$\begin{aligned} P(Y=1,T<t|\mathrm{do}(X=x)) &= \textstyle\sum_z \sum_{x'} P(Y=1,T<t|X=x',Z=z)P(Z=z|X=x)P(X=x') \\ &\simeq \textstyle\sum_z \sum_{x'} e^{\eta_{x'}} e^{\eta_z} H_0(t) P(Z=z|X=x)P(X=x') \\ &= H_0(t)\left(\textstyle\sum_z e^{\eta_z} P(Z=z|X=x)\right)\left(\textstyle\sum_{x'} e^{\eta_{x'}} P(X=x')\right) \end{aligned} \tag{9}$$

Next consider the specific example where $\eta_{x'} = \beta_x x'$, $\eta_z = \beta_z z$, $P(X = x)$ is a normal distribution $N(\mu_x, \sigma_x^2)$, and $P(Z = z|X = x)$ is a normal distribution $N(\alpha x, \sigma_z^2)$, where in the latter case α is a constant and the mean of z is αx. Understanding that the sums should be considered as integrals when variables are continuous, then we have,

$$\sum_{x'} e^{\eta_{x'}} P(X=x') = \exp\left(\beta_x \mu_x\right) \exp\left(\frac{\sigma_x^2 \beta_x^2}{2}\right) \tag{10}$$

and,

$$\sum_{z} e^{\eta_z} P(Z=z|X=x) = \exp\left(\beta_z \alpha x\right) \exp\left(\frac{\sigma_z^2 \beta_z^2}{2}\right) \tag{11}$$

giving,

$$P(Y=1,T<t|\mathrm{do}(X=x)) \simeq H_0(t) \exp(\beta_x \mu_x) \exp\left(\frac{\sigma_x^2\beta_x^2}{2} + \frac{\sigma_z^2\beta_z^2}{2}\right) \exp(\beta_z \alpha x) \tag{12}$$

The incidence rate at baseline $X = x_0$ is determined by the first three terms, and differs from a proportional hazard estimate that is adjusted by either or both, of x or z. The first two terms are equivalent to a proportional hazards estimate with x at the mean exposure μ_x and z at the baseline value, and the third term quantitatively accounts for the spread in values of x and z about their mean values. The influence of $\mathrm{do}(X = x)$, is seen in the last term $e^{\beta_z \alpha x}$, with the change in risk being mediated by z in a very simple and intuitive way.

For the situation considered here, where there is solely an indirect effect of the exposure through the mediator, this estimate is the same as for a mediation analysis with *measured* confounding [9]. Interestingly, in the equivalent mediation analysis with measured confounding, the influence of measured confounding on the estimate[2], does not appear in the resulting expressions for natural direct, and indirect, effects. This appears to explain the agreement between estimates with measured, and unmeasured confounding - for the model of figure 2 in limit of rare diseases and a proportional hazards model, the estimate is (apparently) unaffected by confounding.

Equation 9 applies to any situation described by figure 2, and the example given can be generalised, e.g. to multivariate normal distributions.

[2]For a solely indirect effect, $\gamma_1 = \gamma_3 = 0$ in Eq. 4.6 on page 101 of [9], and measured confounding is accounted for through the coefficient γ_4, that does not subsequently appear in the equations for natural direct and indirect effects.

Attributable fractions

Attributable fractions are intended to describe the proportion of disease incidence that is caused by an exposure, or can be avoided by an intervention. They can be defined in several related but distinct ways [2, 5]. Here the attributable fraction for the situation described by figure 1 is considered. To allow exploration of the causal influence of a subset X, of risk factors X and W, the risk factors are considered to be composed of both X and W. If $P(Y=1,T<t)$ is the probability of observing a disease at a time T, less than t, then the average causal effect of risk factors on disease risk in a population compared with baseline risk factors is, $P(Y=1,T<t)-P(Y=1,T<t|\text{do}(X=x_0))$, and the excess fraction is,

$$A_f = \frac{P(Y=1,T<t)-P(Y=1,T<t|\text{do}(X=x_0))}{P(Y=1,T<t)} \tag{13}$$

The numerator of 13 is the average causal effect (ACE) [6] of the risk factors X on the population's disease risk, compared with the baseline values $X=x_0$. It is divided by the probability of risk in the population, giving an excess risk fraction, that is referred to here as an attributable fraction. To evaluate this, firstly note that,

$$\begin{aligned} P(Y=1,T<t) &= \int dxdwdz\; P(Y=1,T<t,X=x,W=w,Z=z) \\ &= \int dxdwdz\; P(Y=1,T<t|X=x,W=w,Z=z)P(X=x,W=w,Z=z) \\ &\simeq H_0(t)\int dxdwdz\; e^{\eta_x+\eta_w+\eta_z}P(X=x,W=w,Z=z) \end{aligned} \tag{14}$$

where we assumed the data could be described by a proportional hazards model for sufficiently rare diseases, to allow the approximation $F(t)\simeq H(t)=H_0(t)e^{\eta_x+\eta_w+\eta_z}$, where η_x,η_w,η_z are linear predictors respectively involving x, w, and z. Integrals should be replaced by sums for non-continuous variables. $P(Y=1,T<t|\text{do}(X=x_0))$ can be evaluated similarly, and for the example of figure 1 considered here, we can use the backdoor adjustment formula in the second line below,

$$\begin{aligned} P(Y=1,T<t|\text{do}(X=x_0)) &= \int dwdz\; P(Y=1,T<t,W=w,Z=z|\text{do}(X=x_0)) \\ &= \int dwdz\; P(Y=1,T<t|X=x_0,W=w,Z=z)P(W=w,Z=z) \\ &\simeq H_0(t)e^{\eta_{x_0}}\int dwdz\; e^{\eta_w+\eta_z}P(W=w,Z=z) \end{aligned} \tag{15}$$

Therefore, using Eqs. 14 and 15, the excess fraction is given by,

$$A_f = \frac{\int dxdwdz\; e^{\eta_x+\eta_w+\eta_z}P(X=x,W=w,Z=z) - e^{\eta_{x_0}}\int dxdwdz\; e^{\eta_w+\eta_z}P(W=w,Z=z)}{\int dxdwdz\; e^{\eta_x+\eta_w+\eta_z}P(X=x,W=w,Z=z)} \tag{16}$$

where $e^{\eta_{x_0}}=1$ for the baseline variables x_0. The equation is very similar to conventional expressions for attributable fractions that use relative risks [5], that would give an attributable fraction of $1-1/R=1-e^{-\eta_x}$, where R is the relative risk. However the expression now involves averages over the population, that include potential correlations with the confounding variables. The WHO uses an attributable fraction that is defined as [1],

$$A_W = \frac{\int dx\; e^{\eta_x}P(X=x) - \int dx\; e^{\eta_x}P'(X=x)}{\int dx\; e^{\eta_x}P(X=x)} \tag{17}$$

where $P'(X=x)$ is an alternative probability distribution for X. If we take $P'(X=x)$ to be a delta function centred on $X=x_0$, with $e^{x_0}=1$, then,

$$A_W = 1-\frac{1}{\int dx\; e^{\eta_x}P(X=x)} \tag{18}$$

which is the same as would be obtained by assuming that e^{η_x} and $e^{\eta_w+\eta_z}$ are uncorrelated in Eq. 16. Appendix B shows that the A_W provides a lower (upper) bound on A_f if e^{η_x} is positively (negatively) correlated with $e^{\eta_w+\eta_z}$. In general Eqs. 16 and 17 will differ, and neither should have a causal interpretation unless the causal model satisfies suitable conditions such as those in figure 1 that ensure that causal associations are being estimated.

To estimate the integrals in 16, note that $E[f(X)]=E[(1/n)\sum_{i=1}^n f(X_i)]$ and that the variance $\text{Var}[(1/n)\sum_{i=1}^n f(X_i)]=(1/n)\text{Var}(f(X))\to 0$ as $n\to\infty$. This allows the integrals to be approximated by a sum over the observed data, which is reasonable if the number of data points is sufficiently large in each level of categorical data considered. For example, in the

study of UK Biobank described later with nearly 500,000 participants, the smallest category was for current smokers, but this included over 50,000 smokers. With this approximation,

$$A_f \simeq 1 - \frac{\sum_{i=1}^{n} e^{\eta_{w_i}+\eta_{z_i}}}{\sum_{i=1}^{n} e^{\eta_{x_i}+\eta_{w_i}+\eta_{z_i}}} \tag{19}$$

That might alternately be written as,

$$A_f = 1 - \frac{1}{\sum_{i=1}^{n} w_i e^{\eta_{x_i}}} \tag{20}$$

with,

$$w_i = \frac{e^{\eta_{w_i}+\eta_{z_i}}}{\sum_{i=1}^{n} e^{\eta_{w_i}+\eta_{z_i}}} \tag{21}$$

which shows that the relative risk is weighted by the influence of confounders and other risk factors, but is similar to conventional expressions attributable fractions with $A = 1 - e^{-\eta_{x_i}}$. When there are no confounders or other risk factors than x, then the terms in Eq. 21 become 1, and $w_i = 1/n$, so that $\sum_{i=1}^{n} w_i e^{\eta_{x_i}}$ is then just the average of $e^{\eta_{x_i}}$ across the population. The expression makes it clear that if the relative risk $e^{-\eta_{x_i}}$ is positively correlated with the relative risks from the confounding and other potential risk factors $e^{\eta_{w_i}+\eta_{z_i}}$, then $\sum_{i=1}^{n} w_i e^{\eta_{x_i}} > \sum_{i=1}^{n} e^{\eta_{x_i}}$, and the attribution fraction is greater when accounting for the confounding and other potential risk factors.

To compare the attributable risk between setting $X = x_1$ and $X = x_2$, the equivalent expression to Eq. 19 is,

$$\frac{P(Y=1,T<t|\text{do}(X=x_2)) - P(Y=1,T<t|\text{do}(X=x_1))}{P(Y=1,T<t|\text{do}(X=x_2))} = 1 - \frac{e^{\eta_{x_1}} \sum_{i=1}^{n} e^{\eta_{w_i}+\eta_{z_i}}}{e^{\eta_{x_2}} \sum_{i=1}^{n} e^{\eta_{w_i}+\eta_{z_i}}} = 1 - e^{\eta_{x_1}-\eta_{x_2}} \tag{22}$$

which is just the conventional result for attributable fraction in terms of the relative risk.

Number of attributed cases

The proportion of disease cases that are attributed to a risk factor is only important if the disease is sufficiently common. The change in the number of cases of disease can be estimated using the estimated attributable fraction and the number of observed cases of disease. If N is the population size under consideration, and $P \equiv P(Y = 1, T < t)$, $P_0 \equiv P(Y = 1, T < t|\text{do}(X = x_0))$, then,

$$A_f = \frac{N(P - P_0)}{NP} \tag{23}$$

If we approximate NP as the observed number of cases in the population being studied N_{obs}, then we can estimate the number of extra (or fewer) cases from the attributable fraction A_f, with,

$$N_{A_f} \equiv N(P - P_0) = NP \simeq A_f N_{obs} \tag{24}$$

This gives a simple estimate for the number of cases that are attributable to a risk factor. However, this is the number of attributable cases of hospital admissions, for diseases included by the study's selection criterion - first admissions in an ICD-10 chapter in this paper. This latter estimate could substantially differ from our perception of the number of hospital admissions caused by a specific disease, that could be dominated by sequences of hospital visits, or result from a different original underlying cause. For that reason, attributable fractions are generally a better measure of the causal influence of risk factors on the risk of disease.

If the attributable fraction given by Eq. 19 were negative, then instead of considering $(P - P_0)/P$, an alternative would be to consider $(P_0 - P)/P_0$. However, provided A_f is reasonably small, then the two estimates have approximately the same magnitude, with a change in sign to indicate the direction of effect. Expanding $(P_0 - P)/P_0$ in terms of $A_f = (P - P_0)/P$, gives,

$$\frac{P_0 - P}{P_0} = -\left(\frac{P - P_0}{P}\right) \frac{1}{1 - \left(\frac{P-P_0}{P}\right)} \simeq -\left(\frac{P - P_0}{P}\right)(1 + A_f) \tag{25}$$

Showing that both expressions are approximately equal in magnitude if A_f is small.

Effect of treatment on the treated (ETT)

An alternative attribution fraction, that is of more interest to clinicians or an individual, is the chance of having avoided a disease if you had not been exposed, but were subjected to the same correlated confounding factors that you would have otherwise experienced. This situation is equivalent to estimating the "effect of treatment on the treated" (ETT) [6, 7], but the "treatment" is an exposure to smoking or BMI. For the situation considered here of figure 1, this counterfactual question can be formulated and expressed in terms of observational quantities in a similar way to before. The argument below considers the simpler situation of smokers versus never smoked, or max BMI tertile versus a lower BMI tertile, denoting exposed by $X = x_1$ and unexposed by $X = x_0$. Using counterfactual notation where Y_{x_1} indicates the disease status of (e.g.) smokers, and Y_{x_0} the disease status of non-smokers, then the ETT is defined as [6, 7],

$$ETT = [Y_{x_1} - Y_{x_0}|X = x_1] \tag{26}$$

that can be thought of as estimating the difference between disease risk in smokers and non-smokers, when subjected to the same correlated confounding influences as smokers would experience. Following a previous derivation [7], and incorporating the same proportional hazards assumptions as before, this can be written as,

$$\begin{aligned} ETT &= P(Y_{x_1} = 1|X = x_1) - P(Y_{x_0} = 1|X = x_1) \\ &= \int P(Y = 1|W = w, Z = z, X = x_1)P(W = w, Z = z|X = x_1)dwdz \\ &\quad - \int P(Y = 1|W = w, Z = z, X = x_0)P(W = w, Z = z|X = x_1)dwdz \\ &= H_0(t)e^{\eta_{x_1}} \int e^{\eta_w + \eta_z} P(W = w, Z = z|X = x_1)dwdz - H_0(t)e^{\eta_{x_0}} \int e^{\eta_w + \eta_z} P(W = w, Z = z|X = x_1)dwdz \end{aligned} \tag{27}$$

where the second term on the second line is usually justified with the backdoor adjustment formula Eq. 1, but corresponds to estimating the probability of disease when $X = x_0$ but all other exposures are as they would have been if $X = x_1$, and the third line uses the approximation of sufficiently rare diseases that the cumulative distribution function can be approximated by the cumulative hazard. Continuing to take the baseline value $e^{\eta_{x_0}} = 1$, and dividing by the first term to get an attribution fraction, then gives,

$$A_{ETT} = \frac{e^{\eta_{x_1}} - 1}{e^{\eta_{x_1}}} \tag{28}$$

which solely involves the relative risk $R = e^{\eta_{x_1}}$ for e.g. smoking status, and is the simplest attribution fraction that occurs in the literature.

Because survival analyses are designed to estimate the influence of risk on an individual, with hindsight, perhaps Eq. 28 should not have been a surprise? Within the proportional hazards model, smoking will modify your risk of disease, independently of whether any other factors also do. From a population perspective, disease risk is determined by the overall combination of exposures, that will usually be correlated. This is why the attribution fraction for the population needs a more careful estimation that accounts for correlations between the exposures and confounding variables.

Attribution of disease incidence to smoking and BMI

A key requirement for public health policy, is to assess the impact of exposures such as smoking and BMI on health. Depending on the disease, the influence of smoking or BMI can be very different. The associations can change in both magnitude and direction of effect, being harmful for most diseases but protective for some. Their influence is further complicated by the correlation of risk factors with known confounders such as socioeconomic status and education, something that Eq. 19 is intended to account for. To explore the overall influence of smoking and BMI on disease, the attribution of smoking and BMI to the incidence of over 400 hospital diagnosed diseases in men and women was considered using UK Biobank data [10]. The aim was to characterise diseases by their attribution fractions, allowing them to be ranked or classified in terms of their risk modifiability in terms of smoking and BMI. The selection of diseases for study is detailed elsewhere [13], along with further information on the UK Biobank data that was used [10, 13]. Although the survival analyses could be done better by studying each disease in detail and formulating the analysis appropriately, the study here accounts for the strongest confounding factors, while allowing a broad survey of the overall influence of smoking and BMI on disease. The causal model was assumed to be as in figure 1, and the analysis is briefly summarised below.

Survival analysis

The International Classification of Diseases version 10 (ICD-10) [14], classifies disease diagnoses into chapters of related diseases. To minimise the potential for confounding by prior disease, that might modify incidence rates, only the first incidence of disease in each ICD-10 chapter was considered for each individual who lived long enough to experience it. We considered disease diagnoses that were the primary cause of hospital admission and would have passed a threshold of severity to trigger hospital admission, as identified from hospital episode statistics (HES), and recorded with an ICD-10 code. Individuals who reported diabetes at entry to the study were excluded. This ensured that new cases of diabetes would almost entirely involve type II diabetes, with risk most strongly influenced by age and BMI. For each disease studied, data were excluded if the disease occurred before a participant entered the study, or if an individual had a hospital diagnosed cancer other than non-melanoma skin cancer before starting the study. The disease incidence rates are "rare" in the approximate sense needed to estimate attribution fractions [12]. A survival analysis using age as the time variable was left truncated at a participant's entry to the study, right-censored if there was: death, cancer other than non-melanoma skin cancer, or the study period ended. All diagnoses recorded between entering the study and 31st January 2020 were included, as recorded in UK Biobank HES data on 8th December 2021. Data beyond 31st January 2021 were likely to be influenced by the COVID-19 pandemic and were omitted. Analyses were multiply adjusted using a proportional hazards model, with men and women studied separately, and a causal model assumed as in figure 1. Adjustment considered the established risk factors of: smoking status (never, previous, or current), alcohol consumption (rarely - less than 3 times per month, sometimes - less than 3 times a week but more than 3 per month, regularly - 3 or more times each week), education (degree level, post-16 but below degree, to age 16 or unspecified), socio-economic status (tertiles), height (sex-specific tertiles), BMI (sex-specific tertiles), and for women we also adjusted for: HRT use ever (yes, no), and one or more children (yes,no). Baseline was taken as: never smoker, rarely drink, brisk walking pace, degree-level education, minimum deprivation tertile, minimum height tertile in men (or women), middle BMI tertile in men (or women), and women with no children or HRT use. Only diseases with at least 140 cases were considered. This ensured there were at least 10 cases per parameter to adjust from baseline, even if a parametric e.g. Weibull model with an extra two parameters to fit the baseline hazard function were considered [12]. Sensitivity analyses excluded participants with a broader range of prior diseases, leading to fewer total cases and fewer diseases included in the study. Analyses were multiply adjusted. There were less than 1% missing values, allowing a complete case analysis. Numerical work and plots used R [15], and packages used here included: survival[16] and grr[17].

Attribution fractions for the UK Biobank population were considered for three situations: observed population verus all smoking status as never-smoked, observed population versus middle BMI tertile, and observed population versus never-smoked and middle BMI tertile. The latter case is comparing the correlated exposures of BMI and smoking status in the observed population, to a situation where BMI and smoking are set to their baseline values. Because the baseline BMI tertile is the middle tertile, current smokers could be correlated with either the top or bottom BMI tertile. Frequency of alcohol consumption was adjusted for but not studied, because it is a less precise measure than smoking status or BMI, and it is known to have inconsistent associations with disease risk in different studies.

Results

Plots and tables only included diseases with statistically significant associations for current smoking versus never smoked or maximum versus middle BMI tertile, after an FDR multiple-testing adjustment. Where results involve both smoking and BMI then diseases were included if they are included in either of the smoking-only, or BMI-only results. This left 129 diseases associated with BMI, 153 diseases associated with smoking, and 226 diseases that were associated with either smoking or BMI. To explore the sensitivity of the estimates to the strength of confounding factors, estimates made using Eq. 19 and 18 were compared (figure 5 in the Appendix). As expected, the influences of confounding are more noticeable for smaller attributable fractions, but even in those cases, the estimates rarely differ by more than about 20%. With a handful of exceptions, such as Parkinson's disease (G20), estimates with Eq. 19 were larger than with 18, as would be expected if the influence of smoking and BMI were positively correlated with the influence of the confounding factors in the model.

Figure 3 shows the median attributable fractions for the combined influence of smoking and BMI on the incidence of disease in each ICD-10 chapter, with the width of the bar plots proportional to the number of diseases included in the estimate. Diseases of the respiratory system (X) have the largest median attribution fraction, of about 0.3, closely followed by endocrine, nutritional, and metabolic diseases, that are both almost double the next largest values. Diseases of the skin and subcutaneous tissues (XII) and of the nervous system (VI), both have median attribution fractions near 0.15. Neoplasms

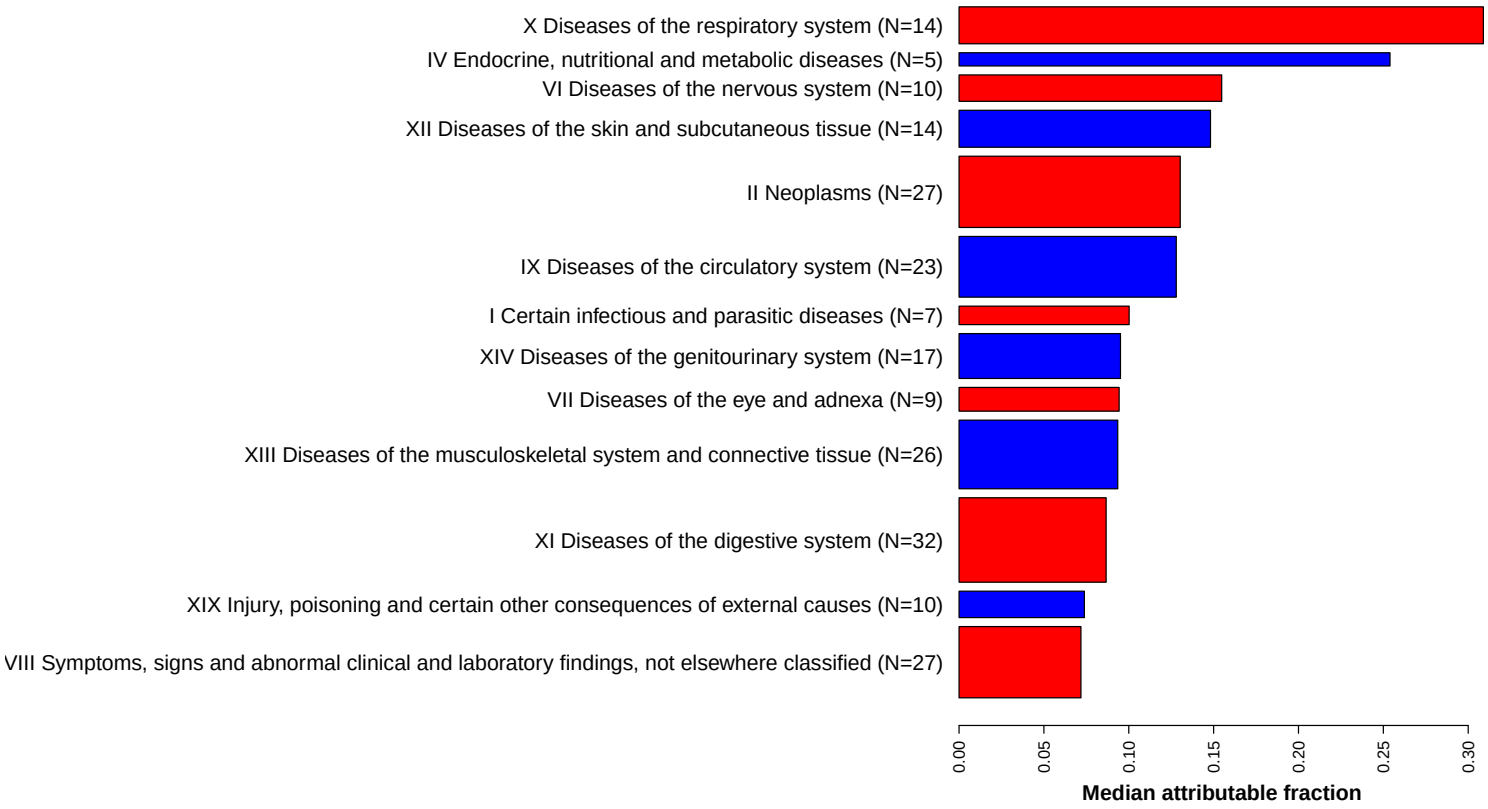


Figure 3: Median attributable fractions for each ICD-10 chapter with at least 5 diseases where $A_f > 0.2$. Bar widths are proportional to the number of diseases in each chapter.

and circulatory diseases account for 22% of all the diseases, and have the next largest median attributable fractions. There are seven chapters with median attribution fractions greater than 0.1, and these include 100 diseases, 50 of which are neoplasms and diseases of the circulatory system.

The 26% of diseases that had $A_f \geq 0.2$ are listed in table 1. There are 11 diseases with $A_f \geq 0.5$ and 21 with $A_f \geq 0.35$. Given the limitations of the analysis and the potential for regression dilution bias, it is possible that more than 11 diseases could have $A_f \geq 0.5$. For diseases with more than half the cases attributed to smoking and BMI, it seems reasonable to regard smoking and BMI as "pathogenic", in a similar way that strong genetic risk factors are often described as pathogenic. One third of the 226 diseases had an attributable fraction with $|A_f| > 0.17$, and two thirds had $|A_f| > 0.06$. Although the mean attribution fraction for the combined influence of smoking and BMI was $\simeq 15\%$, the estimated attributable number of extra cases was only $\simeq 8\%$, reflecting the fact that the most common diseases (with the most cases), tended to have smaller attributable fractions.

Diseases were ranked in terms of their attribution fractions for smoking and BMI (figure 4). Figure 4 identifies an important point, that even established risk factors such as smoking and BMI can have protective associations with some diseases. The 20 diseases that smoking and BMI have the strongest protective associations with are listed in table 2. There were 12 diseases whose protective association had an attributable fraction with magnitude greater than 0.1, and 3 with magnitude greater than 0.2. Melanoma in situ (D03), had the strongest protective association of -0.29, where the sign is taken to indicate the direction of effect as discussed in "Number of attributed cases".

Sensitivity analysis

Participants with hospital reports of prior cancers other than non-melanoma skin cancers were excluded from the main study, but self-reported cancers or other prior diseases were not. It is possible for example, that a heart attack might be followed by weight loss, and including participants with prior heart attacks could weaken a potential association between BMI and

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
E66 Obesity	F	311	306	24	0.98	1
J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified	F	209	193	53	0.92	2
J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection	F	416	376	17	0.90	3
J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection	M	417	374	18	0.90	4
J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified	M	261	230	41	0.88	5
C34 Malignant neoplasm of bronchus and lung	M	1018	833	4	0.82	6
I70 Atherosclerosis	M	156	114	86	0.73	7
C34 Malignant neoplasm of bronchus and lung	F	996	710	6	0.71	8
E11 Non-insulin-dependent diabetes mellitus	M	206	114	86	0.55	9
I71 Aortic aneurysm and dissection	M	402	222	44	0.55	10
R04.2 Haemoptysis	M	314	162	60	0.52	11
C15 Malignant neoplasm of oesophagus	M	473	220	46	0.47	12
R91 Abnormal findings on diagnostic imaging of lung	M	358	162	60	0.45	13
C67 Malignant neoplasm of bladder	M	1063	438	9	0.41	14
R91 Abnormal findings on diagnostic imaging of lung	F	372	151	65	0.41	15
I50 Heart failure	F	280	111	88	0.40	16
K42 Umbilical hernia	F	297	111	88	0.37	17
G47.3 Sleep apnoea	F	381	139	69	0.36	18
J84 Other interstitial pulmonary diseases	F	177	63	117	0.35	19
I50 Heart failure	M	359	125	75	0.35	20
R04.2 Haemoptysis	F	239	83	103	0.35	21
J10 Influenza due to identified influenza virus	F	211	69	114	0.33	22
M13 Other arthritis	M	215	70	112	0.32	23
R29.6 Tendency to fall, not elsewhere classified	M	182	58	123	0.32	24
A41 Other septicaemia	F	957	303	25	0.32	25
J18 Pneumonia, organism unspecified	M	3011	944	2	0.31	26
C22 Malignant neoplasm of liver and intrahepatic bile ducts	M	171	54	126	0.31	27
L72.0 Epidermal cyst	M	456	139	69	0.30	28
J18 Pneumonia, organism unspecified	F	2777	845	3	0.30	29
G47.3 Sleep apnoea	M	776	236	39	0.30	30
N17 Acute renal failure	F	333	99	94	0.30	31
G44.2 Tension-type headache	F	152	44	135	0.29	32
K43 Ventral hernia	F	377	108	90	0.29	33
G62 Other polyneuropathies	M	175	50	130	0.29	34
C16 Malignant neoplasm of stomach	M	260	70	112	0.27	35
B37 Candidiasis	M	173	47	132	0.27	36
R06.0 Dyspnoea	M	650	176	58	0.27	37
I26 Pulmonary embolism	F	836	225	43	0.27	38
R63.4 Abnormal weight loss	F	485	125	75	0.26	39
J84 Other interstitial pulmonary diseases	M	234	60	119	0.26	40
J22 Unspecified acute lower respiratory infection	F	1506	386	15	0.26	41
E87.1 Hypo-osmolality and hyponatraemia	M	234	59	120	0.25	42
I25.9 Chronic ischaemic heart disease, unspecified	M	254	63	117	0.25	43
M48 Other spondylopathies	F	573	139	69	0.24	44
K62.1 Rectal polyp	M	1143	275	32	0.24	45
N17 Acute renal failure	M	562	135	71	0.24	46
H02.4 Ptosis of eyelid	M	253	59	120	0.23	47
C90 Multiple myeloma and malignant plasma cell neoplasms	F	247	58	123	0.23	48
M47 Spondylosis	M	338	79	105	0.23	49
L60 Nail disorders	F	236	55	125	0.23	50
R13 Dysphagia	M	959	218	47	0.23	51
J90 Pleural effusion, not elsewhere classified	M	524	119	82	0.23	52
J22 Unspecified acute lower respiratory infection	M	1349	300	27	0.22	53
L03 Cellulitis	M	2036	450	8	0.22	54
M81 Osteoporosis without pathological fracture	F	946	204	50	0.22	55
K92.0 Haematemesis	M	156	33	152	0.21	56
L03 Cellulitis	F	1602	333	21	0.21	57
K25 Gastric ulcer	M	338	70	112	0.21	58
I64 Stroke, not specified as haemorrhage or infarction	M	173	35	148	0.20	59

Table 1: Attributable fractions A_f for both smoking and BMI are estimated with Eq. 19, ranked, and listed if $A_f \geq 0.2$. Colours: $A_f \geq 0.5$ (red), $0.5 > A_f \geq 0.35$ (orange), $0.35 > A_f \geq 0.2$ (yellow). Sex: diseases in males (M) or females (F), N: total cases, N_{A_f}: attributed cases.

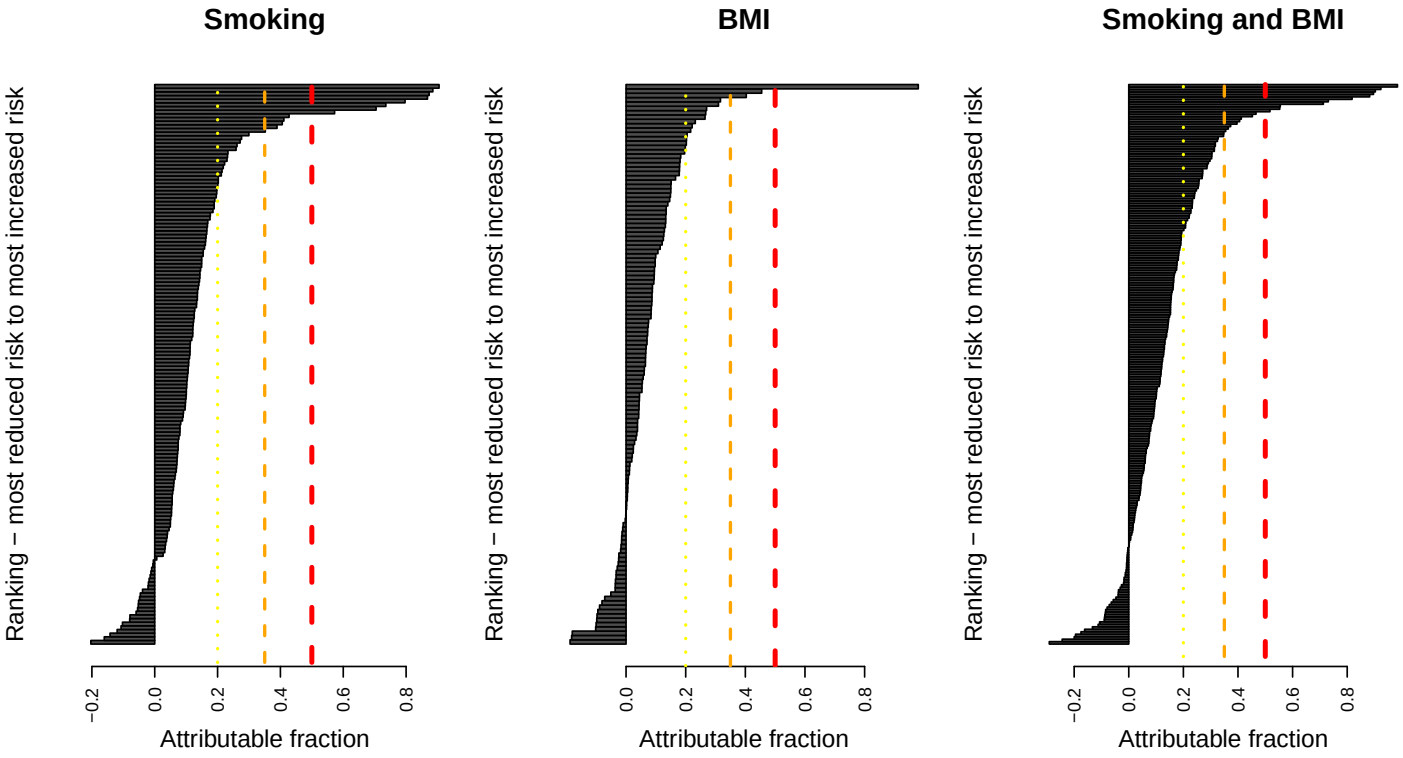


Figure 4: For diseases with a statistically significant association with smoking or BMI after an FDR multiple-testing adjustment, attributable fractions A_f were calculated with Eq. 19. Lines indicate $A_f = 0.5$ (red), $A_f = 0.35$ (orange), $A_f = 0.2$ (yellow). $A_f < 0$ indicates a protective association.

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
D03 Melanoma in situ	M	272	−79	14	−0.290	1
K31.7 Polyp of stomach and duodenum	F	870	−212	9	−0.240	2
N41 Inflammatory diseases of prostate	M	572	−114	10	−0.200	3
N81 Female genital prolapse	F	4199	−819	1	−0.190	4
R79 Other abnormal findings of blood chemistry	M	1905	−333	5	−0.170	5
S02 Fracture of skull and facial bones	M	355	−57	18	−0.160	6
C43 Malignant melanoma of skin	M	723	−96	11	−0.130	7
S76.1 Injury of quadriceps muscle and tendon	M	215	−24	25	−0.110	8
M20.1 Hallux valgus (acquired)	F	2875	−307	6	−0.110	9
C61 Malignant neoplasm of prostate	M	5800	−521	2	−0.090	10
B34 Viral infection of unspecified site	M	319	−28	23	−0.089	11
N40 Hyperplasia of prostate	M	3928	−344	4	−0.088	12
M16 Coxarthrosis [arthrosis of hip]	M	3167	−272	7	−0.086	13
J90 Pleural effusion, not elsewhere classified	F	327	−28	23	−0.084	14
K40 Inguinal hernia	F	493	−39	20	−0.079	15
R19.8 Other specified symptoms and signs involving the digestive system and abdomen	F	302	−22	27	−0.072	16
C44 Other malignant neoplasms of skin	M	6095	−391	3	−0.064	17
K31.7 Polyp of stomach and duodenum	M	300	−17	31	−0.057	18

Table 2: Diseases with the strongest protective associations, ranked by the proportion of disease attributed to a combination of smoking and BMI (A_f). Sex indicates diseases in males (M) or females (F), N are total cases, N_{A_f} are the number of cases attributed to smoking and BMI, A_f is the attributable fraction.

heart disease. In contrast, smoking might increase the risk of some diseases for which a substantial proportion occur before entry into the UK Biobank study. In that case, including participants with the prior disease might strengthen the associations. The question of how best to study sequences of disease is an example where causal understanding is not enough, and new statistical methods or data are likely to be required. It might be an intractable question, due to the vast possible combinations

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
S00.8 Superficial injury of other parts of head	F	171	39	135	0.23	44
S92 Fracture of foot, except ankle	M	168	38	138	0.22	46
E21 Hyperparathyroidism and other disorders of parathyroid gland	F	296	66	106	0.22	48
I21 Acute myocardial infarction	F	1125	243	28	0.22	50
M79.6 Pain in limb	F	920	198	41	0.22	51
M17 Gonarthrosis [arthrosis of knee]	F	3623	735	3	0.20	57

Table 3: The sensitivity analyses found six additional diseases with $A_f \geq 0.2$, for the combination of both smoking and BMI, that would have appeared in table 1. Sex: diseases in males (M) or females (F), N: total cases, N_{A_f}: attributed cases.

of sequences of 100s of diseases, and it is further complicated by the complex time-dependent exposures and accumulation of genetic mutations that any individual experiences. Therefore a senstivity analysis compared the paper's main results with a second analysis that excluded participants who had reported any cancer other than non-melanoma skin cancer, or any serious cardiovascular disease of heart disease, stroke, arterial or pulmonary embolisms, or subarachnoid haemorrhage.

Differences between the main study and the sensitivity analysis were small (for full details see the Supplementary Material). The difference in attribution fractions between the two studies had a mean and median of -0.006 and -0.005 respectively, and a standard deviation of 0.029. The differences in magnitude were typically equivalent to about 10%. The attribution fractions of six diseases changed by more than 0.05. These included increased attributable fractions for: I50 - heart failure in women (0.40 to 0.48), R29.6 - tendency to fall in men (0.32 to 0.41), and decreases in: C16 - stomach cancer in men (0.27 to 0.21), J10 - influenza in women (0.33 to 0.27), J22 - lower respiratory infections in men (0.24 to 0.17), and N17 - acute renal failure (0.24 to 0.17). Overall the differences were small, but larger than the approximations in Eq. 19 would cause.

Discussion

Attribution fractions

Similar attribution formulae to that used here have existed in published literature since at least 1998 [4]. A contribution of this paper is to recognise that for a particular causal model such as that in figure 1, the attribution fractions can only be used with a restricted range of potential risk modifiers, whose associations have a causal interpretation. The consideration of attributable fractions from the perspective of causal inference, has helped to highlight several important points:

1. It is not possible to estimate causal associations and attributable fractions for all factors that are thought to be associated with disease risk.

For example, if the causal model is not known sufficiently well, then it may not be possible to correctly adjust for confounding. Alternately, if the measurement is too imprecise e.g. socio-economic status is likely to capture the influence of several factors that may include exposure to pollution, poor quality diet, poor living and working conditions, etc, then it may not be possible to estimate a meaningful causal association - for example, someone with an equivalent socio-economic status in a different country would experience different exposures and have different causal factors that influence their health.

2. Estimates of causal associations cannot always be obtained directly from a single multiply-adjusted analysis, and may need several different analyses of the same dataset, as required by the causal model.

For example, changes in systolic blood pressure (SBP) can be caused by smoking or BMI, and therefore SBP should not be adjusted for if we are interested in the influence of smoking and BMI on disease risk. In contrast, if our interest was in SBP, then we would need to adjust for BMI and smoking if they can modify disease risk in any way other than through changes in SBP.

Attribution to smoking and BMI

The original intention of this work was to characterise diseases through the overall modifiability of disease risk by established risk factors. To do so as rigorously as possible, a causal model similar to figure 1 was assumed, and the strongest established risk factors of smoking and BMI were considered. Under the model in figure 1, Eq. 19 is intended to provide attribution fractions for the causal increase in disease incidence due to smoking and BMI, while accounting for confounding factors that would also be expected to modify disease incidence. The study considered the most common diseases in UK Biobank, and highlighted two important points:

1. Attribution fractions for smoking, and BMI, are very heterogeneous, and can even involve a reduction in risk (table 2).
2. Some associations are extremely strong, for example with $A_f > 0.5$.

The first point highlights a difficulty in optimising lifestyle and drug treatments - changes in lifestyle or medication are likely to have a heterogeneous influence on disease risk, with some risks lowered but others potentially increased. In the same way that some germline genetic risk factors are described as pathogenic when they substantially increase your risk of disease, it seems reasonable to describe the influence of smoking and BMI as "pathogenic", for diseases with high attributable fractions such as $A_f > 0.5$. Based on the estimates here, eliminating smoking and controlling BMI would be expected to prevent the majority of those diseases in an equivalent population.

The attribution fractions allow us to identify the diseases for which a change in lifestyle is likely to have the greatest impact. From a population perspective, eliminating smoking and controlling BMI in an equivalent population would be expected to avoid: the majority of diseases with $A_f > 0.5$ (red in table 1), between one third and one half of diseases with $0.35 > A_f > 0.5$ (orange in table 1), between one fifth and one third of diseases with $0.2 > A_f > 0.35$ (yellow in table 1). This slightly ad-hoc categorisation of diseases by their attributable fractions provides an approximate indication of how the patterns of disease would be expected to change if smoking were eliminated and BMI were controlled in a population that was otherwise similar to that in UK Biobank.

The attributed fractions would be larger if more of the population were exposed. For examples where a binary exposure X is uncorrelated with W or confounders Z, then Appendix C shows that provided $p(R-1) \ll 1$, where R is the relative risk and p is the proportion of the population that are exposed, then $A_f \simeq p(R-1)$. In that case, if the exposed proportion p were halved, then so would the attributable fraction. This highlights a limitation of Eqs. 16, 19, and 17 - they measure the proportion of disease in a population that is attributed to an exposure, but a clinician might be more interested in what proportion of disease in smokers is attributable to smoking, and an individual might be more interested in whether their risk of serious disease or death is substantially changed. Such questions that refer to individuals as opposed to populations might best be tackled with counterfactual arguments and measurements such as Eq. 28. An alternative approach is to consider the "probability of necessity" [6, 7], that in principle allows the assessment of whether it is more probable than not that the disease would not have occurred if you had not been exposed to e.g. smoking. Such approaches allow specific individual cases to be assessed, but do not provide an overall characterisation of an exposure's influence on population health.

When considering the attribution fraction for both smoking and BMI together, the study included diseases with statistically significant associations with either smoking or BMI. In this situation, especially when the number of cases are few, the estimates for one of the two parameters can in principle be both large and imprecise. This has the potential to produce misleading estimates for the joint attribution fraction of both smoking and BMI. A notable example is the strong protective association of smoking with Parkinson's disease (see table 7), that was substantially weakened in table 2 by the association with BMI, even though the association with BMI was not statistically significant. This appears to be an isolated example, and the potential problem will reduce with more cases, but it highlights the importance of also considering the attribution fractions associated with each separate exposure.

Meta analyses

If estimates are to be used in a meta-analyses of data that have not originated from a randomised control trial, then it is essential to ensure that estimates are of causal associations. It is possible that some reported estimates from observational studies will measure the causal influence of a potential risk factor, such as BMI, alcohol, and smoking in the first example

considered here, but this is unlikely to be true for all variables that are adjusted for. If a study has inappropriately adjusted for potential confounding variables, then the data cannot be included in the meta-analysis. The assessment of this requires a good causal understanding of how the risk factor of interest modifies disease risk, and the potential confounding factors that need adjusting for. Disagreement between studies may indicate incomplete understanding of the underlying causal model, with inappropriate or insufficient adjustment for confounding factors. In the common situation where uncertainty of the causal processes linking exposure X to disease risk remain, then the standard methods and cautious reporting of conventional epidemiology must remain [2, 8].

Conclusions

The aim was to characterise how the risk of common diseases are modified by established risk factors, and to identify which disease risks are the most sensitive to them. This requires estimates of causal associations between potential risk factors and disease risk, and it was shown that many conventional epidemiological studies can provide these. This requires a sufficiently good model of the causal processes, and appropriate adjustment for confounders. This limits the range of potential risk factors that can be studied, but will often include important risk factors such as smoking and BMI.

The first generic example was a causal model with risk factors X that satisfied the "backdoor criteria" [7], a situation that is likely to commonly occur. The second example considered the "frontdoor" criteria, that can allow causal estimations in the presence of unmeasured confounders, when the influence of X is mediated by a measurable variable Z. In this case, for the simple example considered (with rare disease incidence, and the linear influence of a continuous exposure mediated by a normally distributed continuous variable whose mean is linearly related to the exposure), the causal estimate is the same as you get from a mediation analysis with measured confounders. With hindsight, this might have been anticipated by observing that for this situation in the limit of rare diseases studied with a proportional hazards model, the confounding terms do not appear in the results of mediation studies for natural direct, and indirect, effects [9].

The causal attribution of exposures to population disease risk was considered for the situation corresponding to figure 1, leading to equations 16 and 19. These are similar to the expression used by the World Health Organisation, but accounts for correlations between exposures and confounding factors. The attribution fractions (A_f) allowed a simple categorisation of diseases to indicate how the patterns of disease might change if smoking were eliminated and BMI controlled, in an equivalent population to UK Biobank. For the combined influence of smoking and BMI compared to baseline of never smoking and mid-tertile BMI, this identified 11 diseases with $A_f > 0.5$, a further 10 with $0.35 < A_f < 0.5$, and a total of 59 with $A_f > 0.2$. Similar results were found by a sensitivity analysis. Because a large proportion of the population studied were "unexposed" (non-smoking and a healthy weight), the attribution fraction for e.g. disease in the population attributed to smoking, will often be less than that for an individual. Attribution fractions for individuals are simpler expressions (Eq. 28), because they do not need to account for complex correlations that occur between risk factors in a population.

In summary, this article has attempted to link established epidemiological methods [2, 3] with newer techniques in causal inference [6, 7], to help clarify the definition of attribution fractions. This subsequently allowed a quantitative characterisation of the causal associations between smoking, BMI, and the incidence of disease in the UK Biobank cohort [10], using their attribution fractions (A_f), and categorising them into groups with similar A_f. These help to indicate the expected change in patterns of disease due to changes e.g. to a healthier lifestyle, for an equivalent population.

Data availability

UK Biobank data can be accessed by application through www.ukbiobank.ac.uk, and summary data produced during this study will become available from:

`https://osf.io/`

UK Biobank has approval by the Research Ethics Committee (REC) under approval number 16/NW/0274. UK Biobank obtained participant's consent for the data to be used for health-related research.

Code availability

R code used to produce figures from summary data will become available with the summary data at:

`https://osf.io/`

The full code for use with non-summary data will be returned with other results to UK Biobank (see www.ukbiobank.ac.uk).

Acknowledgements

Thank you to Professor Robert Clarke for suggesting a sensitivity analysis to strengthen the results. This research has been conducted using the UK Biobank resource under application number 42583. Anthony Webster is supported by an intermediate research fellowship from the Nuffield Department of Population Health (NDPH), University of Oxford.

Competing interests

The author declares no competing interests.

References

[1] World Health Organization, *Global health risks: mortality and burden of disease attributable to selected major risks*, World Health Organization (2009).

[2] T.L. Lash, T.J. VanderWeele, S. Haneuse, K.J. Rothman, *Modern Epidemiology*, Fourth Edition, Wolters Kluwer, (2021).

[3] D. Collett *Modelling Survival Data in Medical Research*, New York: Chapman and Hall/CRC, 3rd edition, (2014).

[4] Rockhill, B. Newman, B. and Weinberg, C. *Use and misuse of population attributable fractions* American Journal of Public Health, **88**, 15-19, (1998).

[5] M.A. Mansournia, A. Douglas G, *Population attributable fraction*, BMJ - British Medical Journal, **360**, k757, (2018).

[6] J. Pearl *Causality*, 2nd ed., John Wilely & Sons Ltd, (2009).

[7] J. Pearl, M. Glymour, N.P. Jewell, *Causal Inference In Statistics*, Cambridge University Press, (2016).

[8] Shimonovich, M., Pearce, A., Thomson, H. et al. *Assessing causality in epidemiology: revisiting Bradford Hill to incorporate developments in causal thinking* Eur. J. Epidemiol. **36**, 873-887 (2021).

[9] T.J. VanderWeele *Explanation in Causal Inference*, Oxford University Press, (2015).

[10] C. Bycroft et al. *The UK Biobank resource with deep phenotyping and genomic data* Nature **562**, 203-209, (2018).

[11] E.T. Jaynes *Probability Theory: The Logic of Science*, Cambridge University Press, (2003).

[12] A.J. Webster, R. Clarke *Sporadic, late-onset, and multistage diseases*, medRxiv 2021.12.15.21267843, Cold Spring Harbor Laboratory Press, (2021).

[13] Webster, A.J., Gaitskell, K., Turnbull, I., Cairns B.J., Clarke R. *Characterisation, identification, clustering, and classification of disease* Scientific Reports **11**, 5405 (2021).

[14] World Health Organisation, International Statistical Classification of Diseases (ICD), www.who.int/standards/classifications/classification-of-diseases, (2021).

[15] R Core Team. *R: A Language and Environment for Statistical Computing* R Foundation for Statistical Computing, Vienna, Austria, www.R-project.org, (2020).

[16] Therneau, T.M. A Package for Survival Analysis in R, CRAN.R-project.org/package=survival, (2021).

[17] Varrichio, C. *grr: Alternative Implementations of Base R Functions*, CRAN.R-project.org/package=grr, (2016).

[18] Yang, L. Kartsonaki, C, Yao, P. *The relative and attributable risks of cardia and non-cardia gastric cancer associated with Helicobacter pylori infection in China: a case-cohort study*, The Lancet Public Health, **6**, Issue 12, e888 - e896.

A Probability densities and hazard functions

Using Eq. 4 we can calculate the probability density function,

$$\begin{aligned} P(Y=1, T \in (t, t+\delta t)|\text{do}(X=x)) &= \delta t \tfrac{d}{dt} P(Y=1, T<t|\text{do}(X=x)) \\ &\simeq e^{\eta_x} A_Z h_0(t) \delta t \\ &= e^{\eta_x} A_Z P(Y=1, T \in (t, t+\delta t)|X=x_0, Z=z_0) \end{aligned} \tag{29}$$

This allows a hazard function (usually defined as $h = f/S = f/(1-F)$), to be defined as,

$$h(t|\text{do}(X=x))\delta t = \frac{P(Y=1, T \in (t, t+\delta t)|\text{do}(X))}{1 - P(Y=1, T<t|\text{do}(X))} \simeq e^{\eta_x} A_Z h_0(t) \delta t \tag{30}$$

Alternately, we could have argued that most diseases in UK Biobank are sufficiently rare that we can approximate $f(t) \simeq h(t)$, and hence that $h(t|\text{do}(X=x))\delta t \simeq P(Y=1, T \in (t, t+\delta t)|\text{do}(X=x))$ Therefore for combinations of confounders and risk factors that satisfy the "backdoor criteria" [7] (such as those in figure 1), and disease incidence that is sufficiently rare (which includes most studies of the first incidence of a disease in UK Biobank data [12]),

$$\frac{h(t|\text{do}(X=x))}{h(t|\text{do}(X=x_0))} \simeq e^{\eta_x} \tag{31}$$

This indicates that if we are interested in the relative difference in disease rates that would be caused by changing from the baseline value $X = x_0$, to $X = x$ (e.g. to reduce risk), then the result is given in terms of the usual relative risk, but without terms in Z (corresponding to $Z = z_0$). This indicates that causal inferences for x using the relative risk, will be correct in these circumstances. Any confounders $Z \neq z_0$ that may appear in an estimated relative risk, are set at their baseline values $Z = z_0$ to estimate Eqs. 31 and 6.

Note that hazard functions do not give a probability in the conventional sense, and are not normalised to 1 for example. To ensure the above calculations were done correctly and will e.g. normalise to 1 (within the limits of the approximations made), it was essential for the arguments to have used probabilities (for which the backdoor theorem applies). For a clear exposition of when a particular situation will satisfy the "backdoor criterion", please refer to the references [6, 7].

B Bounds on attribution fractions

By definition, if the relative risks e^{η_x} and $e^{\eta_w+\eta_z}$ are positively correlated, then $E[e^{\eta_x} e^{\eta_w+\eta_z}] - E[e^{\eta_x}]E[e^{\eta_w+\eta_z}] \geq 0$, where E is used to denote expectations. Therefore, using $\exp(s) \geq 0$ for any real-valued s, we can rearrange the inequality as,

$$\begin{aligned} E[e^{\eta_x} e^{\eta_w+\eta_z}] &\geq E[e^{\eta_x}]E[e^{\eta_w+\eta_z}] \\ \tfrac{1}{E[e^{\eta_x}]} &\geq \tfrac{E[e^{\eta_w+\eta_z}]}{E[e^{\eta_x} e^{\eta_w+\eta_z}]} \\ 1 - \tfrac{E[e^{\eta_w+\eta_z}]}{E[e^{\eta_x} e^{\eta_w+\eta_z}]} &\geq 1 - \tfrac{1}{E[e^{\eta_x}]} \end{aligned} \tag{32}$$

Eq. 32 shows that if the relative risks for x, w, and z are positively correlated, then the attributable fraction for disease risk within the population, is greater than would be estimated using the average relative risk, with estimates using the mean relative risk providing a lower bound. If x and z are negatively correlated then the $\geq$ sign is be replaced by $\leq$.

Another quantity that might be considered is the expected value of the attributed fraction $1-1/e^{\eta_x}$, that is $E[1-1/e^{\eta_x}] = 1 - E[1/e^{\eta_x}]$. Because $1/e^{\eta_x}$ is concave, Jensen's inequality gives,

$$E\left[\frac{1}{e^{\eta_x}}\right] \geq \frac{1}{E[e^{\eta_x}]} \tag{33}$$

and as a result,

$$1 - \frac{1}{E[e^{\eta_x}]} \geq 1 - E\left[\frac{1}{e^{\eta_x}}\right] = E\left[1 - \frac{1}{e^{\eta_x}}\right] \tag{34}$$

If e^{η_x} and $e^{\eta_w+\eta_z}$ are positively correlated, then Eqs. 34 and 32 indicate that $E[1-1/e^{\eta_x}]$ will also bound Eq. 19. However, this would not be true if e^{η_x} and $e^{\eta_w+\eta_z}$ were negatively correlated.

C Relation to other attributable fractions

A recent study [18] with a proportion p exposed to a virus, and an estimated relative risk R, reported an attribution fraction of $A = p(R-1)/R$. Here it is briefly outlined when this will approximate Eq. 19. Assume that e^{η_x}, e^{η_w}, and e^{η_z} are uncorrelated, so that Eq. 16 simplifies to 17, that may be approximated as,

$$A_f \simeq 1 - \frac{1}{\frac{1}{n}\sum_{i=1}^{n} e^{\eta_{x_i}}} \tag{35}$$

If we consider a proportion p that are exposed with relative risk R, and a proportion $(1-p)$ that are unexposed, then using Eq. 35,

$$\begin{aligned} A_f &\simeq 1 - \tfrac{1}{(1-p)+pR} \\ &= \tfrac{p(R-1)}{1+p(R-1)} \\ &= p\left(1-\tfrac{1}{R}\right)\left(\tfrac{R}{1+p(R-1)}\right) \\ &\simeq p\left(1-\tfrac{1}{R}\right) \end{aligned} \tag{36}$$

where the approximation in the final line follows if $R-1$ is small enough, as it often can be. A better approximation follows from the second line, where $p(R-1) \ll 1$ ensures that $A_f \simeq p(R-1)$. If $p \simeq 1$, then $A_f \simeq (R-1)/R$ as usual, as can be seen from the first or second line above.

D Supplementary tables and figures

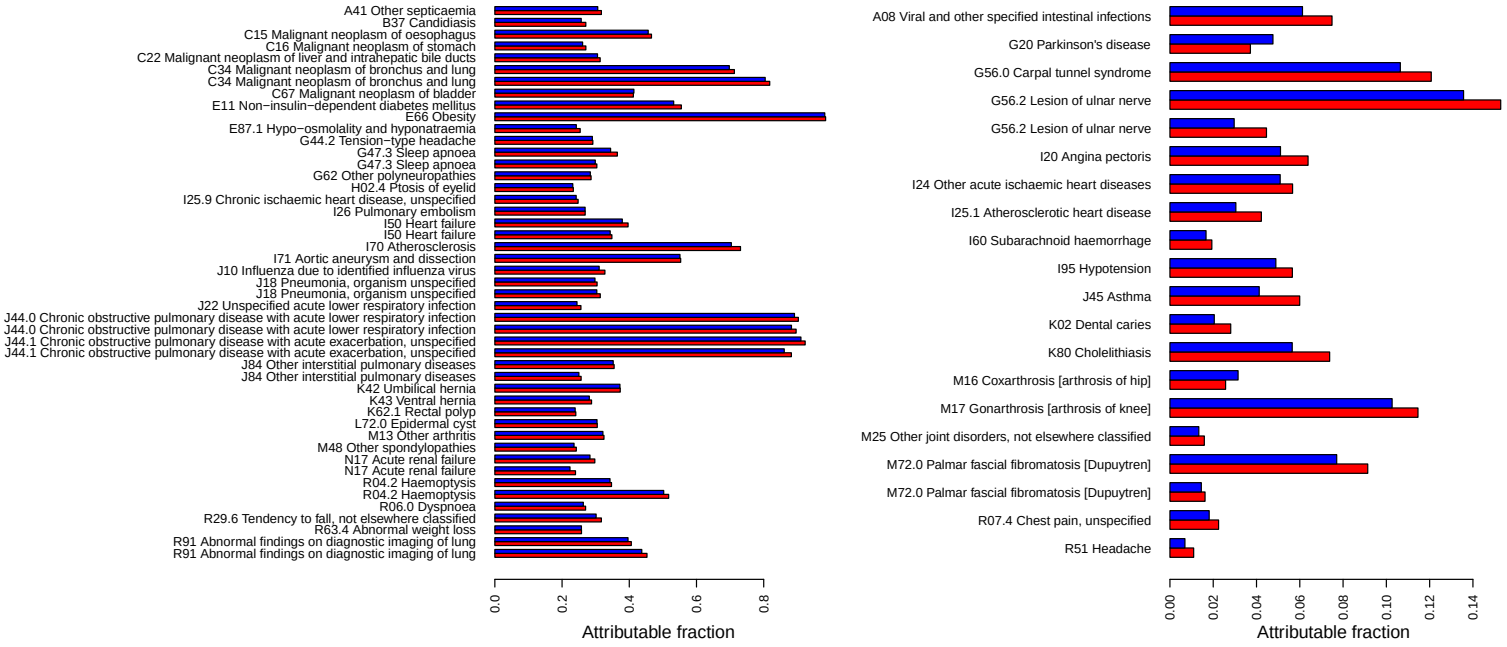


Figure 5: Attributable fractions (A_f) for smoking and BMI, estimated with Eqs. 19 (red) and 18 (blue). Eq. 18 is equivalent to the WHO's A_f [1]. Eq. 19 estimates causal associations. Left plot: diseases with the highest 25% A_f. Right plot: diseases where Eqs. 18 and 19 differed the most.

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
E66 Obesity	F	311	305	7	0.98	1
E11 Non-insulin-dependent diabetes mellitus	M	206	94	38	0.45	2
K42 Umbilical hernia	F	297	120	32	0.40	3
G47.3 Sleep apnoea	F	381	121	31	0.32	4
I50 Heart failure	M	359	111	34	0.31	5
G62 Other polyneuropathies	M	175	47	66	0.27	6
M13 Other arthritis	M	215	57	53	0.27	7
I50 Heart failure	F	280	74	44	0.26	8
I26 Pulmonary embolism	F	836	194	16	0.23	9
G47.3 Sleep apnoea	M	776	172	21	0.22	10
J22 Unspecified acute lower respiratory infection	F	1506	326	6	0.22	11
J10 Influenza due to identified influenza virus	F	211	43	73	0.20	12
I25.9 Chronic ischaemic heart disease, unspecified	M	254	51	59	0.20	13
M81 Osteoporosis without pathological fracture	F	946	191	17	0.20	14

Table 4: BMI only: Sex indicates diseases in males (M) or females (F), N are total cases, N_{A_f} are cases attributed to BMI, A_f is the attributable fraction for deviations from the mid-tertile of BMI. For obesity, unsurprisingly $A_f \simeq 1$. Reporting errors may have prevented $A_f = 1$ for obesity.

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified	F	209	189	28	0.90	1
J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection	F	416	368	10	0.89	2
J44.0 Chronic obstructive pulmonary disease with acute lower respiratory infection	M	417	364	12	0.87	3
J44.1 Chronic obstructive pulmonary disease with acute exacerbation, unspecified	M	261	227	21	0.87	4
C34 Malignant neoplasm of bronchus and lung	M	1018	811	2	0.80	5
I70 Atherosclerosis	M	156	115	50	0.74	6
C34 Malignant neoplasm of bronchus and lung	F	996	702	3	0.70	7
I71 Aortic aneurysm and dissection	M	402	230	20	0.57	8
R04.2 Haemoptysis	M	314	134	43	0.43	9
C15 Malignant neoplasm of oesophagus	M	473	194	26	0.41	10
R91 Abnormal findings on diagnostic imaging of lung	M	358	145	37	0.41	11
C67 Malignant neoplasm of bladder	M	1063	414	8	0.39	12
R91 Abnormal findings on diagnostic imaging of lung	F	372	131	45	0.35	13
J84 Other interstitial pulmonary diseases	F	177	53	83	0.30	14
B37 Candidiasis	M	173	48	90	0.28	15
J84 Other interstitial pulmonary diseases	M	234	64	72	0.27	16
J90 Pleural effusion, not elsewhere classified	M	524	137	41	0.26	17
R04.2 Haemoptysis	F	239	62	75	0.26	18
G56.2 Lesion of ulnar nerve	F	222	52	85	0.23	19
K92.0 Haematemesis	M	156	36	102	0.23	20
C22 Malignant neoplasm of liver and intrahepatic bile ducts	M	171	39	97	0.23	21
R29.6 Tendency to fall, not elsewhere classified	M	182	40	94	0.22	22
R06.0 Dyspnoea	M	650	142	39	0.22	23
J18 Pneumonia, organism unspecified	M	3011	645	4	0.21	24
C16 Malignant neoplasm of stomach	M	260	55	79	0.21	25
H02.0 Entropion and trichiasis of eyelid	M	337	68	69	0.20	26
K62.1 Rectal polyp	M	1143	231	18	0.20	27
G44.2 Tension-type headache	F	152	31	110	0.20	28

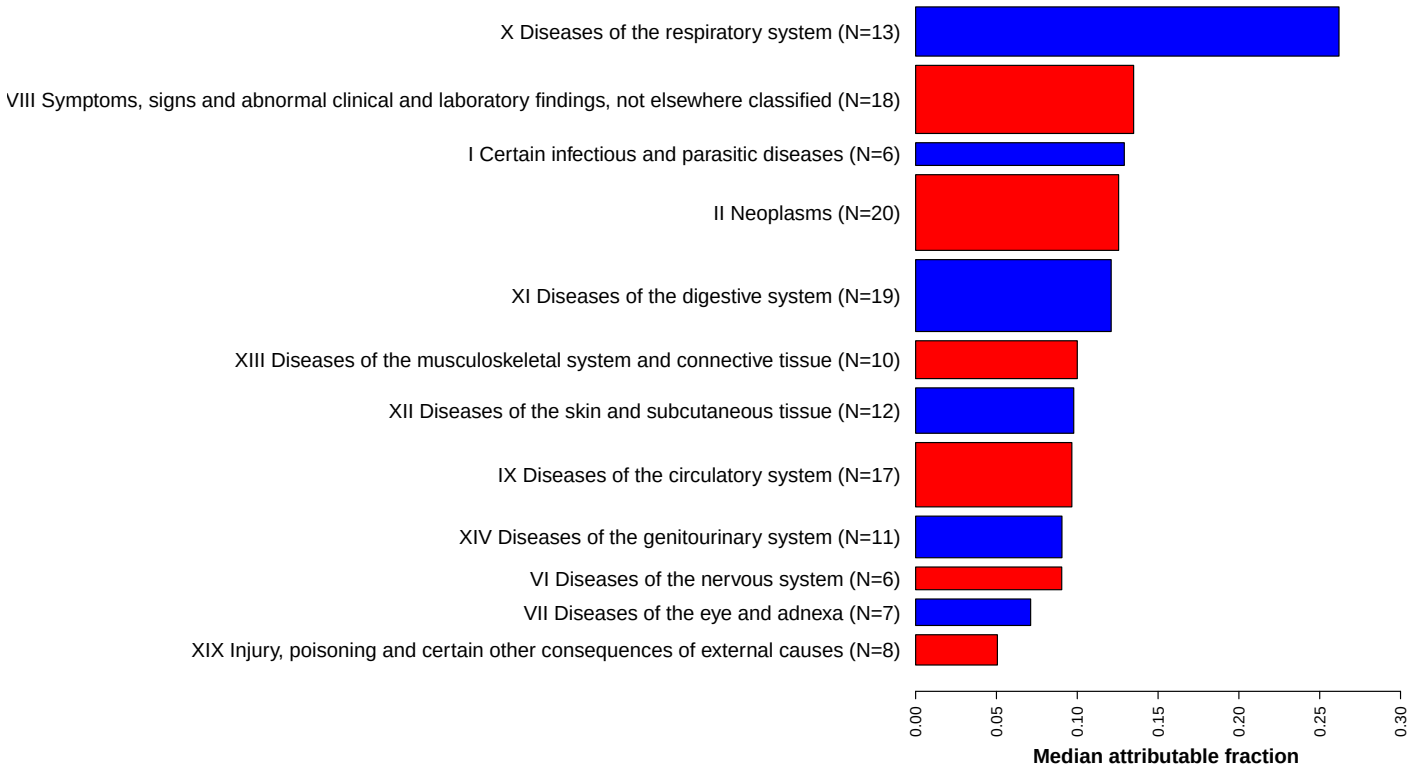
Table 5: Smoking only: Sex indicates diseases in males (M) or females (F), N are total cases, N_{A_f} are cases attributed to smoking, A_f is the attributable fraction.

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
N81 Female genital prolapse	F	4199	−788	1	−0.190	1
L72.0 Epidermal cyst	F	525	−96	13	−0.180	2
N41 Inflammatory diseases of prostate	M	572	−103	12	−0.180	3
S02 Fracture of skull and facial bones	M	355	−36	23	−0.100	4
K40 Inguinal hernia	F	493	−49	20	−0.100	5
R19.8 Other specified symptoms and signs involving the digestive system and abdomen	F	302	−30	25	−0.098	6
M20.1 Hallux valgus (acquired)	F	2875	−280	2	−0.097	7
S76.1 Injury of quadriceps muscle and tendon	M	215	−20	26	−0.095	8
R79 Other abnormal findings of blood chemistry	M	1905	−169	5	−0.089	9
D04 Carcinoma in situ of skin	M	201	−16	27	−0.081	10
K52.9 Non-infective gastro-enteritis and colitis, unspecified	M	1061	−76	15	−0.071	11
S82 Fracture of lower leg, including ankle	F	2163	−110	10	−0.051	12

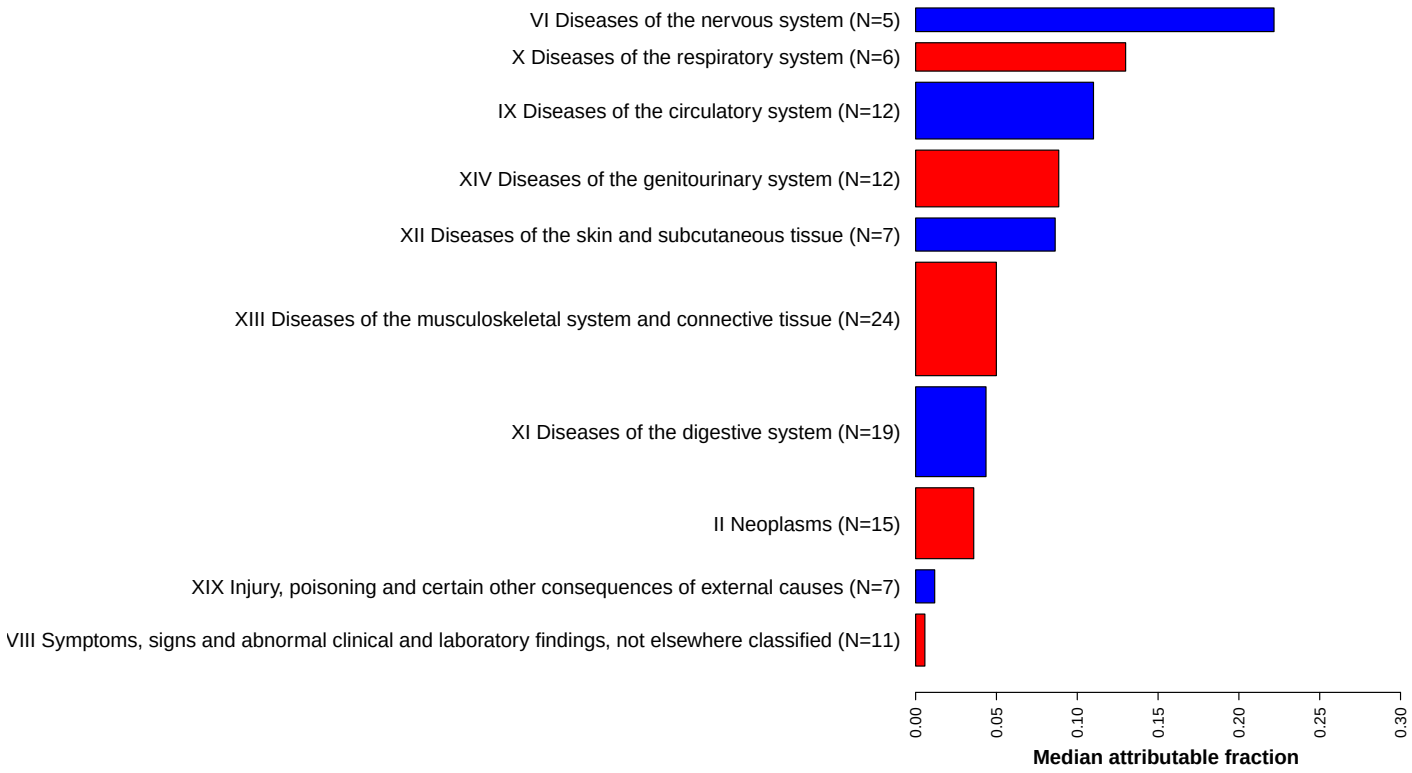
Table 6: BMI only: Diseases with the strongest protective associations, ranked by attributable fraction (A_f). Sex indicates diseases in males (M) or females (F), N are total cases, N_{A_f} are the number of cases attributed to BMI.

Disease	Sex	N	N_{A_f}	Rank	A_f	Rank A_f
G20 Parkinson's disease	F	152	−31	18	−0.200	1
D03 Melanoma in situ	M	272	−44	14	−0.160	2
G20 Parkinson's disease	M	289	−41	15	−0.140	3
C43 Malignant melanoma of skin	M	723	−86	8	−0.120	4
S52 Fracture of forearm	M	687	−74	9	−0.110	5
K31.7 Polyp of stomach and duodenum	F	870	−89	7	−0.100	6
C54 Malignant neoplasm of corpus uteri	F	878	−70	12	−0.080	7
R79 Other abnormal findings of blood chemistry	M	1905	−150	4	−0.079	8
R19.5 Other fecal abnormalities	F	544	−32	17	−0.059	9
N40 Hyperplasia of prostate	M	3928	−214	3	−0.055	10
C61 Malignant neoplasm of prostate	M	5800	−304	1	−0.052	11
K31.7 Polyp of stomach and duodenum	M	300	−15	21	−0.051	12

Table 7: Smoking only: Diseases with the strongest protective associations, ranked by attributable fraction (A_f). Sex indicates diseases in males (M) or females (F), N are total cases, N_{A_f} are the number of cases attributed to smoking.



Attribution fractions for smoking (existing UK Biobank population, versus if all never smoked)



Attribution fractions for BMI (existing UK Biobank population, versus if all were mid-tertile BMI)

Figure 6: Median attributable fractions for each ICD-10 chapter with at least 5 diseases where $A_f > 0.2$. Bar widths are proportional to the number of diseases in each chapter.