

## Supplementary File 1: OncoArray Genotyping

Genomic DNA samples from the 23 lung cancer studies and 11 head and neck cancer studies were genotyped at the Center for Inherited Disease Research, the Helmholtz Center Munich, Copenhagen University Hospital, and the University of Cambridge, using the custom Illumina OncoArray. The quality control strategy for OncoArray data has been described in detail elsewhere (1).

Genotype assignment was performed blinded to case control status for 57775 individuals and 533,631 single nucleotide polymorphisms (SNPs) using the Illumina Genome Studio algorithm. This included 44591 samples (43398 individuals and 1193 QC duplicate samples) part of the lung cancer component of the OncoArray collaboration, 12901 individuals from other unrelated OncoArray studies, and 283 HapMap control individuals of European, African, Chinese and Japanese origin. After removing expected duplicates, standard quality control procedures and OncoArray consortium filters were applied to exclude underperforming DNA samples and genotyping assays (1). Individuals with high missingness ( $>10\%$ ) and low call rate ( $<95\%$ ) were excluded. Genotype data for X chromosomes was used to identify and exclude samples for which reported and chromosomal sex were discordant. Poorly genotyped SNPs (call rate  $<95\%$ ) and variants for which the genotype distributions deviated from Hardy-Weinberg equilibrium ( $p < 10^{-7}$  in controls or  $p < 10^{-12}$  in cases) were also removed (1). Identity by descent (IBD) analysis was used to identify unexpected duplicates ( $IBD > 0.95$ ) and suspected relatives with IBD values between 0.45 and 0.95.

The final 5p15.33 region for the TL association analysis extended from 1250 to 1530 kb, and contained 1446 variants from 4 genes (*TERT*, *CLPTM1L*, *SLC6A3* and *LPCAT1*), which were included on the OncoArray based on a targeted deep sequencing of 288 lung cancer case-control pairs over a 250 kb region in 5p15.33 (2). Data were filtered to remove monomorphic variants ( $n=547$ ), leaving a total of 899 variants for analysis in 2051 individuals, with an overall genotyping rate of 0.998.

## References

1. Amos CI, Dennis J, Wang Z, et al. The OncoArray Consortium: a Network for Understanding the Genetic Architecture of Common Cancers. *Cancer Epidemiol Biomarkers Prev* 2016.
2. Zuzarte PC, Denroche RE, Fehringer G, et al. A two-dimensional pooling strategy for rare variant detection on next-generation sequencing platforms. *PLoS One* 2014;9(4):e93455.

## **Supplementary File 2: Measurement of telomere length**

### *Toronto Mount Sinai and Princess Margaret Hospital (MSH-PMH) Study*

Relative TL was measured using a standard protocol performed with 7900HT Real Time system (Life Technologies) (1). This method expresses TL as a ratio (T/S) of telomere repeat length (T) to copy number of a single copy 36B4 gene (S) in each sample. Two PCR reactions were performed for each sample, the first to determine the cycle threshold (Ct) value for telomere amplification and the second to assess the Ct value for the single copy gene amplification. A 5-point standard curve was generated in each reaction plate, consisting of a reference genomic DNA diluted in 5-fold series from 110 ng to 0.176 ng.

Thermal cycling profile of the telomere amplification consisted of initial denaturation at 95°C for 10 min, followed by 35 cycles of 94°C for 15s, 56°C for 1:30 min, and 72°C for 15s. The thermal profile for the single copy gene (36B4) amplification was 95°C for 10 min, followed by 36 cycles of 94°C for 15 s, 60°C for 1 min. qPCR was performed in a 15 µL reaction containing 1x Power SYBR Green PCR Master Mix (Life Technologies), 20 ng (2µL) DNA. The same reference DNA was included in each plate to be used as a calibrator sample to normalize data and standardize across reactions.

All DNA samples were run in triplicate for telomere and single-copy gene reactions, and the mean of 3 cycle threshold values was calculated. The coefficient of determination ( $R^2$ ) for each standard curve was  $\geq 0.99$ . Assay precision was examined by calculating the coefficient of variation (CV) for the Ct values across triplicate qPCR reactions, and the intra-class correlations (ICC) for the resulting T/S ratios. For the telomere reaction, the CV was 0.33% and for the single-copy gene reaction the CV was 0.23%. The ICC for the triplicate T/S ratios was 0.949 (95% CI: 0.945 - 0.951). The maximum CV for both reactions was <2%. The CV in Ct values across plates was 3.87% for the telomere reaction and 1.14% for the 36B4 reaction.

### *Copenhagen General Population Study (CGPS)*

For CGPS participants, absolute TL was measured on a CFX384 real-time PCR detection system (Bio-Rad Laboratories, Denmark), using Monochrome Multiplex qPCR (2), as

described previously (3, 4). The single-copy albumin gene was amplified simultaneously in the same well as the telomere sample. Each reaction was run in quadruplicates and the mean of 4 cycle threshold values was calculated. To determine the absolute TL in base pairs, the length of the calibrator relative to the reference DNA was measured using Southern Blot. The base-pair change per 1.0 T/S unit was subsequently used to calculate absolute TL.

## References

1. Cawthon RM. Telomere measurement by quantitative PCR. *Nucleic Acids Res* 2002;30(10):e47.
2. Cawthon RM. Telomere length measurement by a novel monochrome multiplex quantitative PCR method. *Nucleic Acids Res* 2009;37(3):e21.
3. Weischer M, Bojesen SE, Cawthon RM, et al. Short telomere length, myocardial infarction, ischemic heart disease, and early death. *Arterioscler Thromb Vasc Biol* 2012;32(3):822-9.
4. Rode L, Nordestgaard BG, Bojesen SE. Long telomeres and cancer risk among 95 568 individuals from the general population. *International journal of epidemiology* 2016;45(5):1634-43.

### Supplementary File 3: Selection of 5p15.33 lung cancer susceptibility variants

Chromosome 5p15.33 is an established lung cancer susceptibility region. The aim of the present analysis was to prioritize risk variants for the mediation analysis by identifying susceptibility loci that capture the main lung cancer association signal in the 5p15.33 region. Logistic regression was used to assess the association between 899 variants in 5p15.33 and risk of lung cancer overall (16396 cases) and lung adenocarcinoma (5690 cases) using data from the previously described 23 OncoArray studies (Supplementary Table S1). The main effect of each variant was tested using unconditional logistic regression, assuming a log-additive genetic model. Models were adjusted for age, sex, study, cigarette pack-years, and 10 genetic ancestry principal components.

To determine the significance threshold for identifying the preliminary set of candidate variants the simpleM algorithm<sup>1</sup> was used to calculate the effective number of independent tests ( $M_{\text{eff}}$ ), which was subsequently applied into the Bonferroni correction formula. The multiple testing adjusted significance level was set at  $P < 8.0 \times 10^{-5}$  based on  $M_{\text{eff}} = 625$ . Variants meeting this significance threshold were further filtered to retain independent loci ( $r^2 < 0.20$ ).

A total of 60 variants were significantly associated with lung cancer risk at the multiple-testing corrected threshold of  $P < 8.0 \times 10^{-5}$ . After LD-pruning five independent loci were retained. The top-ranking susceptibility variant was rs421629 ( $P = 1.2 \times 10^{-16}$ ), a common SNP (MAF=0.42) in *CLPTM1L*, which was previously associated with lung cancer<sup>2-4</sup> and has been shown to influence methylation of the *TERT* promoter in lung tumour tissues<sup>5</sup>. Four additional, common risk variants were identified in *TERT*: rs2736108 ( $P = 1.8 \times 10^{-11}$ ), rs56345976 ( $P = 3.6 \times 10^{-9}$ ), rs7705526 ( $P = 8.0 \times 10^{-7}$ ) and rs13167280 ( $P = 1.1 \times 10^{-6}$ ). Of the 56 variants significantly associated with lung adenocarcinoma, the same variants emerged as the top-ranking, independent susceptibility loci: rs7705526 ( $P = 4.6 \times 10^{-13}$ ), rs2736108 ( $P = 1.7 \times 10^{-12}$ ), rs421629 ( $P = 6.2 \times 10^{-9}$ ), rs13167280 ( $P = 1.4 \times 10^{-8}$ ), and rs56345976 ( $P = 2.2 \times 10^{-7}$ ). All of these variants represent previously reported susceptibility loci for lung cancer overall and lung adenocarcinoma<sup>2-4, 6</sup>. Two of the *TERT* susceptibility variants, rs7705526 and rs2736108, were previously identified by Bojesen et al.<sup>7</sup> as the SNPs anchoring two independent association peaks in 5p15.33 for leukocyte telomere length.

For the purposes of the mediation analysis, the total effects of the selected 5p15.33 susceptibility variants were re-estimated after removing overlapping subjects with the Zhang et al.<sup>8</sup> Mendelian randomization analysis, which provided the estimates of the effect of telomere length on lung cancer risk (mediator-outcome relationship). There were no overlapping subjects between our analysis and Bojesen et al.<sup>7</sup>, which provided the estimates of the association between the selected 5p15.33 lung cancer risk variants and leukocyte telomere length (exposure-mediator relationship).

## References

1. Gao X, Starmer J, Martin ER. A multiple testing correction method for genetic association studies using correlated single nucleotide polymorphisms. *Genetic epidemiology* 2008; **32**: 361-9.
2. McKay JD, Hung RJ, Gaborieau V, et al. Lung cancer susceptibility locus at 5p15.33. *Nat Genet* 2008; **40**: 1404-6.
3. Pande M, Spitz MR, Wu X, Gorlov IP, Chen WV, Amos CI. Novel genetic variants in the chromosome 5p15.33 region associate with lung cancer risk. *Carcinogenesis* 2011; **32**: 1493-9.
4. McKay JD, Hung RJ, Han Y, et al. Large-scale association analysis identifies new lung cancer susceptibility loci and heterogeneity in genetic susceptibility across histological subtypes. *Nat Genet* 2017; **49**: 1126-32.
5. Scherf DB, Sarkisyan N, Jacobsson H, et al. Epigenetic screen identifies genotype-specific promoter DNA methylation and oncogenic potential of CHRNA4. *Oncogene* 2013; **32**: 3329-38.
6. Kachuri L, Amos CI, McKay JD, et al. Fine mapping of chromosome 5p15.33 based on a targeted deep sequencing and high density genotyping identifies novel lung cancer susceptibility loci. *Carcinogenesis* 2016; **37**: 96-105.
7. Bojesen SE, Pooley KA, Johnatty SE, et al. Multiple independent variants at the TERT locus are associated with telomere length and risks of breast and ovarian cancer. *Nat Genet* 2013; **45**: 371-84, 84e1-2.
8. Zhang C, Doherty JA, Burgess S, et al. Genetic determinants of telomere length and risk of common cancers: a Mendelian randomization study. *Human molecular genetics* 2015; **24**: 5356-66.

#### Supplementary File 4: Sensitivity analysis for mediation in the presence of interaction

In the presence of interaction between the exposure and mediator, the mediation analysis would be based on the following outcome model:

$$\text{logit}\{P(Y = 1 | a, m, c)\} = \theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c$$

The corresponding formula for the natural indirect effect mediated through a continuous mediator becomes:

$$\log\left(\text{OR}_c^{\text{NIE}}\right) \approx (\theta_2 \times \beta_1 + \theta_3 \times \beta_1 a)(a - a^*)$$

with the formula for natural direct effect  $\log(\text{OR}_c^{\text{NDE}})$  now having a more complicated form, given by Valeri & VanderWeele 2013, p. 150. If the mediation analysis is based on external estimates for the mediator-outcome relationship, we do not have an estimate for  $\theta_3$  available, but can do a sensitivity analysis under different interaction scenarios, based on estimates from marginal (over  $A$  and over  $M$ ) outcome models  $P(Y = 1|m, c)$  and  $P(Y = 1|a, c)$ . In principle, because mediator-exposure interaction induces some interaction between the mediator and exposure and the covariates  $c$ , the marginal models should allow for this. If the models have been fitted without such interaction terms, we can still approximate mediator and exposure effects evaluated at average covariate levels (zero values of centered covariates) with the average mediator and exposure effects, which we will do below. Throughout, in addition to the usual assumptions required for estimating natural direct and indirect effects, we assume that the outcome is rare so that  $\text{logit}\{P(Y = 1|\dots)\} \approx \log\{P(Y = 1|\dots)\}$ , and that the external estimate for the mediator-outcome relationship is valid (for example, based on Mendelian Randomization with a valid instrument).

Suppose that we have an estimate for the parameter  $\theta_2^*$  in the structural model

$$P(Y_{Am} = 1 | c) = E[Y_{Am} | c] \approx \exp(\theta_0^* + \theta_2^* m + \theta_4'^* c)$$

This model can be expanded as:

$$E[Y_{Am} | c] = E_{A|c}\{E[Y_{Am} | A, c]\} = E_{A|c}[\exp(\theta_0 + \theta_1 A + \theta_2 m + \theta'_4 c)]$$

In the absence of interaction we have:

$$\begin{aligned} E[Y_{Am} | c] &= E_{A|c}[\exp(\theta_0 + \theta_1 A + \theta_2 m + \theta'_4 c)] \\ &= \exp(\theta_0 + \theta_2 m + \theta'_4 c) E_{A|c}[\exp(\theta_1 A)] \end{aligned}$$

Thus, the marginal (over  $A$ ) risk ratio is given by:

$$\exp(\theta_2^*) \approx \frac{E[Y_{Am=1} | c]}{E[Y_{Am=0} | c]} = \frac{\exp(\theta_0 + \theta_2 + \theta'_4 c) E_{A|c}[\exp(\theta_1 A)]}{\exp(\theta_0 + \theta'_4 c) E_{A|c}[\exp(\theta_1 A)]} = \exp(\theta_2)$$

meaning that in the absence of interaction the causal risk ratios are collapsible.

In the presence of exposure-mediator interaction we have

$$E[Y_{Am} | c] = E_A[\exp(\theta_0 + \theta_1 A + \theta_2 m + \theta_3 Am + \theta'_4 c)]$$

which further expands to:

$$\begin{aligned} \exp(\theta_2^*) &\approx \frac{E[Y_{Am=1} | c]}{E[Y_{Am=0} | c]} = \frac{\exp(\theta_0 + \theta_2 + \theta_4 c) E_{A|c}[\exp\{(\theta_1 + \theta_3)A\}]}{\exp(\theta_0 + \theta'_4 c) E_{A|c}[\exp(\theta_1 A)]} \\ &= \frac{E_{A|c}[\exp\{(\theta_1 + \theta_3)A\}]}{E_{A|c}[\exp(\theta_1 A)]} \exp(\theta_2) \end{aligned}$$

It follows that if we have estimate the main effect of the exposure  $\theta_1$  and provide values for the exposure mediator interaction parameter  $\theta_3$ , then we can solve for the main effect of the mediator. Suppose that we also have an estimate for the parameter  $\theta_1^\dagger$  in the structural model:

$$P(Y_{aM_a} = 1 | c) = E[Y_{aM_a} | c] \approx \exp(\theta_0^\dagger + \theta_1^\dagger a + \theta_4^{\dagger'} c)$$

This model can be expanded as:

$$\begin{aligned} E[Y_{aM_a} | c] &= E_{M_a|c}\{E[Y_{aM_a} | M_a, c]\} \\ &= E_{M_a|c}[\exp(\theta_0 + \theta_1 a + \theta_2 M_a + \theta_3 a M_a + \theta'_4 c)] \end{aligned}$$

The marginal (over A) risk ratio is given by:

$$\begin{aligned} \exp(\theta_1^\dagger) &\approx \frac{E[Y_{1M_1} | c]}{E[Y_{0M_0} | c]} \\ &= \frac{E_{M_1|c}[\exp(\theta_0 + \theta_1 + \theta_2 M_1 + \theta_3 M_1 + \theta'_4 c)]}{E_{M_0|c}[\exp(\theta_0 + \theta_2 M_0 + \theta'_4 c)]} \\ &= \frac{E_{M_1|c}[\exp\{(\theta_2 + \theta_3)M_1\}]}{E_{M_0|c}[\exp(\theta_2 M_0)]} \exp(\theta_1) \\ \Leftrightarrow \theta_1 &\approx \theta_1^\dagger + \log \left( \frac{E_{M_0|c}[\exp(\theta_2 M_0)]}{E_{M_1|c}[\exp\{(\theta_2 + \theta_3)M_1\}]} \right) \equiv \theta_1^\dagger + f(\theta_2, \theta_3) \end{aligned}$$

Thus, the main effect of the exposure  $\theta_1$  can be expressed in terms of the total effect of the exposure  $\theta_1^\dagger$ , the mediator main effect,  $\theta_2$  and the exposure-mediator interaction  $\theta_3$ . Further, by combining the two equations we get:

$$\begin{aligned} \theta_2^* &\approx \theta_2 + \log \left( \frac{E_{A|c}[\exp\{(\theta_1 + \theta_3)A\}]}{E_{A|c}[\exp(\theta_1 A)]} \right) \\ &= \theta_2 + \log \left( \frac{E_{A|c}[\exp\{(\theta_1^\dagger + f(\theta_2, \theta_3) + \theta_3)A\}]}{E_{A|c}[\exp\{(\theta_1^\dagger + f(\theta_2, \theta_3))A\}]} \right) \end{aligned}$$



Given an external estimate  $\hat{\theta}_2^*$ , for instance from a Mendelian randomization analysis, and the total exposure effect estimate  $\hat{\theta}_1^\dagger$ , and a fixed value of the exposure-mediator interaction  $\theta_3$ , we can solve the equation numerically to obtain an estimate of the mediator main effect  $\theta_2$ , which will provide all of the necessary inputs for estimated the NIE and NDE.

In order to solve the equation for the mediator main effect, we also need to model  $A|c$  and a model for  $|a, c$ , both estimated in the control group. If the mediator  $M$  is continuous, such as telomere length measured in base pairs or T/S ratio units, a linear regression model can be used, assuming normality in the outcome distribution. For a binary mediator, such as a measure of telomere length operationalized as short vs. long, a logistic model was used.

The NIE risk ratio for increase from the reference level  $a^*$  to index level  $a$  can be expressed as:

$$\begin{aligned} \frac{E[Y_{aM_a} | c]}{E[Y_{a^*M_{a^*}} | c]} &= \frac{E_{M_a|c}\{E[Y_{aM_a} | M_a, c]\}}{E_{M_{a^*}|c}\{E[Y_{a^*M_{a^*}} | M_{a^*}, c]\}} \\ &= \frac{E_{M_a|c}[\exp(\theta_0 + \theta_1 a + \theta_2 M_a + \theta_3 a M_a + \theta'_4 c)]}{E_{M_{a^*}|c}[\exp(\theta_0 + \theta_1 a + \theta_2 M_{a^*} + \theta_3 a M_{a^*} + \theta'_4 c)]} \\ &= \frac{E_{M_a|c}[\exp(\theta_2 M_a + \theta_3 a M_a)]}{E_{M_{a^*}|c}[\exp(\theta_2 M_{a^*} + \theta_3 a M_{a^*})]} \end{aligned}$$

Similarly, the NDE for the risk ratio is given by:

$$\begin{aligned} \frac{E[Y_{aM_{a^*}} | c]}{E[Y_{a^*M_{a^*}} | c]} &= \frac{E_{M_{a^*}|c}\{E[Y_{aM_{a^*}} | M_{a^*}, c]\}}{E_{M_{a^*}|c}\{E[Y_{a^*M_{a^*}} | M_{a^*}, c]\}} \\ &= \frac{E_{M_{a^*}|c}[\exp(\theta_0 + \theta_1 a + \theta_2 M_{a^*} + \theta_3 a M_{a^*} + \theta'_4 c)]}{E_{M_{a^*}|c}[\exp(\theta_0 + \theta_1 a^* + \theta_2 M_{a^*} + \theta_3 a^* M_{a^*} + \theta'_4 c)]} \\ &= \frac{E_{M_{a^*}|c}[\exp(\theta_1 a + \theta_2 M_{a^*} + \theta_3 a M_{a^*})]}{E_{M_{a^*}|c}[\exp(\theta_1 a^* + \theta_2 M_{a^*} + \theta_3 a^* M_{a^*})]} \end{aligned}$$

For example, substituting in the outcome and mediator models

$$P(Y = 1 | a, m, c) \approx \exp\{\theta_0 + \theta_1 a + \theta_2 m + \theta_3 am + \theta'_4 c\}$$

$$P(M = 1 | a, c) = \text{expit}\{\beta_0 + \beta_1 a + \beta'_2 c\},$$

the NIE odds ratio can be approximated by

$$\begin{aligned} \frac{E[Y_{aM_a} | c]}{E[Y_{a^*M_{a^*}} | c]} &= \frac{\exp(\theta_2 + \theta_3 a) \text{expit}\{\beta_0 + \beta_1 a + \beta'_2 c\} + [1 - \text{expit}\{\beta_0 + \beta_1 a + \beta'_2 c\}]}{\exp(\theta_2 + \theta_3 a) \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\} + [1 - \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\}]} \approx \text{OR}_c^{\text{NIE}} \end{aligned}$$

and the NDE odds ratio by

$$\frac{E[Y_{aM_{a^*}} | c]}{E[Y_{a^*M_{a^*}} | c]} = \frac{\exp(\theta_1 a + \theta_2 + \theta_3 a) \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\} + \exp(\theta_1 a) [1 - \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\}]}{\exp(\theta_1 a^* + \theta_2 + \theta_3 a^*) \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\} + \exp(\theta_1 a^*) [1 - \text{expit}\{\beta_0 + \beta_1 a^* + \beta'_2 c\}]} \approx \text{OR}_c^{\text{NDE}},$$

which correspond to the formulas given by Valeri & VanderWeele 2013, p. 150 for the binary outcome and binary mediator case. Formulas without exposure-mediator interaction are obtained simply by setting  $\theta_3 = 0$ .

R code for implementing the sensitivity analyses is provided below for both the binary and continuous mediator settings.

# A simulated demonstration of sensitivity analysis for natural direct and indirect effects in the presence of interaction (binary mediator):

```
library(rms)
library(boot)
library(mvtnorm)
```

```
rm(list=ls())
```

```
expit <- function(x) {1/(1+exp(-x))}
logit <- function(p) {log(p)-log(1-p)}
getmode <- function(x) {
  ux <- unique(x)
  ux[which.max(tabulate(match(x, ux)))]
}
```

```
set.seed(1)
nobs <- 10000
```

```
beta0 <- 0.0
beta1 <- 0.25
beta2 <- 0.25
theta0 <- -3.0
theta1 <- 0.25
theta2 <- 0.25
theta3 <- 0.25
theta4 <- 0.25
alpha0 <- 0.0
alpha1 <- 0.25
```

```
# Study 1 (exposure-outcome and exposure-mediator, observational):
```

```
c1 <- rnorm(nobs)
# pa1 <- expit(alpha0 + alpha1 * c1)
# a1 <- rbinom(nobs, 1, pa1)
```

```

# table(a1)/nobs
a1 <- rnorm(nobs, mean=alpha0 + alpha1 * c1, sd=1)
pm1 <- expit(beta0 + beta1 * a1 + beta2 * c1)
m1 <- rbinom(nobs, 1, pm1)
table(m1)/nobs
py1 <- expit(theta0 + theta1 * a1 + theta2 * m1 + theta3 * a1 * m1 + theta4 * c1)
y1 <- rbinom(nobs, 1, py1)
table(y1)/nobs

# a1 <- a1 - mean(a1)

# Full model (not available):

model1 <- glm(y1 ~ a1 + m1 + a1 * m1 + c1, family=binomial(link=logit))
summary(model1)

# Marginal exposure-outcome model:

y1model <- glm(y1 ~ a1 + c1, family=binomial(link=logit))
summary(y1model)

# Exposure-mediator model:

m1model <- glm(m1 ~ a1 + c1, family=binomial(link=logit))
summary(m1model)

# Exposure model:

a1model <- glm(a1 ~ c1, family=gaussian(link=identity))
summary(a1model)

# Study 2 (mediator-outcome, mediator randomized):

c2 <- rnorm(nobs)
# pa2 <- expit(alpha0 + alpha1 * c2)
# a2 <- rbinom(nobs, 1, pa2)
# table(a2)/nobs
a2 <- rnorm(nobs, mean=alpha0 + alpha1 * c2, sd=1)
pm2 <- expit(beta0 + 0.0 * a2 + 0.0 * c2)
m2 <- rbinom(nobs, 1, pm2)
table(m2)/nobs
py2 <- expit(theta0 + theta1 * a2 + theta2 * m2 + theta3 * a2 * m2 + theta4 * c2)
y2 <- rbinom(nobs, 1, py2)
table(y2)/nobs

# a2 <- a2 - mean(a2)

# Full model (not available):

model2 <- glm(y2 ~ a2 + m2 + a2 * m2 + c2, family=binomial(link=logit))
summary(model2)

# Marginal mediator-outcome model:

```

```
y2model <- glm(y2 ~ m2 + c2, family=binomial(link=logit))
summary(y2model)
```

```
# Calculate theta1 at given level of theta2:
```

```
gettheta1 <- function(theta2val, theta3hat, theta1dag, beta0hat, beta1hat) {
  ma <- function(x, aval) {
    checkfinite <- exp((theta2val + theta3hat * aval) * x)
    lp <- beta0hat + beta1hat * aval
    return(ifelse(is.finite(checkfinite), checkfinite, .Machine$double.xmax) *
      exp(x * lp)/(1.0 + exp(lp)))
  }
  ma0 <- ma(x=0, aval=0) + ma(x=1, aval=0)
  ma1 <- ma(x=0, aval=1) + ma(x=1, aval=1)
  return(theta1dag + log(ma0/ma1))
}
```

```
# Function to calculate difference between theta2star and its estimate (the root of
this is the estimate of theta2):
```

```
getdiff <- function(theta2val, theta3hat, theta1dag, theta2star, beta0hat, beta1hat,
alpha0hat, asd) {
  am <- function(x, mval) {
    checkfinite <- exp((theta1hat + theta3hat * mval) * x)
    return(ifelse(is.finite(checkfinite), checkfinite, .Machine$double.xmax) *
      dnorm(x, mean=alpha0hat, sd=asd))
  }
  results <- rep(NA, length(theta2val))
  for (i in 1:length(theta2val)) {
    theta1hat <- gettheta1(theta2val[i], theta3hat, theta1dag, beta0hat, beta1hat)
    am0 <- integrate(am, lower=-Inf, upper=Inf, mval=0)$value
    am1 <- integrate(am, lower=-Inf, upper=Inf, mval=1)$value
    results[i] <- theta2val[i] + log(am1/am0) - theta2star
  }
  return(results)
}
```

```
# Choose a value for the exposure-mediator interaction (true value used for
demonstration):
```

```
theta3val <- 0.25
```

```
# Inputs from the fitted models:
```

```
alpha0hat <- coef(a1model)[1]
asd <- sigma(a1model)
beta0hat <- coef(m1model)[1]
beta1hat <- coef(m1model)[2]
theta1dag=coef(y1model)[2]
theta2star=coef(y2model)[2]
```

```
# Plot the function:
```

```

grid <- seq(-1, 1, by=0.1)
y <- getdiff(grid, theta3hat=theta3val, theta1dag=theta1dag, theta2star=theta2star,
  beta0hat=beta0hat, beta1hat=beta1hat, alpha0hat=alpha0hat, asd=asd)
plot(grid, y, type='l', xlab=expression(theta[2]), ylab='Difference', ylim=c(-1,1))
abline(h=0.0, lty='dashed')

# Solve for theta2 (main effect of mediator on outcome):

theta2hat <- uniroot(getdiff, interval=c(-1,3), theta3hat=theta3val,
  theta1dag=theta1dag, theta2star=theta2star,
  beta0hat=beta0hat, beta1hat=beta1hat, alpha0hat=alpha0hat,
  asd=asd)$root
abline(v=theta2hat, lty='dashed')
cat('\n', 'Estimated theta2=', round(theta2hat, 3), ' with theta3 fixed at ',
  theta3val, '\n\n', sep='')

# Get the corresponding theta1 (main effect of exposure on outcome):

theta1hat <- gettheta1(theta2hat, theta3val, theta1dag, beta0hat, beta1hat)
cat('\n', 'Estimated theta1=', round(theta1hat, 3), ' with theta3 fixed at ',
  theta3val, '\n\n', sep='')

# Natural direct and indirect effects for index versus reference level of exposure:

aind <- 1
aref <- 0

# Decomposition formulas for NIE and NDE estimates:

NIE <- log(exp(theta2hat + theta3val * aind) * expit(beta0hat + beta1hat * aind) +
  (1.0 - expit(beta0hat + beta1hat * aind))) -
  log(exp(theta2hat + theta3val * aind) * expit(beta0hat + beta1hat * aref) + (1.0
  - expit(beta0hat + beta1hat * aref)))
NIE
NIE.OR<-exp(NIE); NIE.OR

NDE <- log(exp(theta1hat * aind + theta2hat + theta3val * aind) * expit(beta0hat +
  beta1hat * aref) + exp(theta1hat * aind) * (1.0 - expit(beta0hat + beta1hat * aref)))
-
  log(exp(theta1hat * aref + theta2hat + theta3val * aref) * expit(beta0hat +
  beta1hat * aref) + exp(theta1hat * aref) * (1.0 - expit(beta0hat + beta1hat * aref)))
NDE
NDE.OR<-exp(NDE); NDE.OR

# A simulated demonstration of sensitivity analysis for natural direct and indirect
  effects in the presence of interaction (continuous mediator):

rm(list=ls())

expit <- function(x) {1/(1+exp(-x))}
logit <- function(p) {log(p)-log(1-p)}
getmode <- function(x) {

```

```

    ux <- unique(x)
    ux[which.max(tabulate(match(x, ux)))]
  }

set.seed(1)
nobs <- 10000
beta0 <- 0.0
beta1 <- 0.25
beta2 <- 0.25
theta0 <- -3.0
theta1 <- 0.25
theta2 <- 0.25
theta3 <- 0.25
theta4 <- 0.25
alpha0 <- 0.0
alpha1 <- 0.25

# Study 1 (exposure-outcome and exposure-mediator, observational):

c1 <- rnorm(nobs)

a1 <- rnorm(nobs, mean=alpha0 + alpha1 * c1, sd=1)
m1 <- rnorm(nobs, mean=beta0 + beta1 * a1 + beta2 * c1, sd=1)
py1 <- expit(theta0 + theta1 * a1 + theta2 * m1 + theta3 * a1 * m1 + theta4 * c1)
y1 <- rbinom(nobs, 1, py1)
table(y1)/nobs

# Full model (not available):

model1 <- glm(y1 ~ a1 + m1 + a1 * m1 + c1, family=binomial(link=logit))
summary(model1)

# Marginal exposure-outcome model:

y1model <- glm(y1 ~ a1 + c1, family=binomial(link=logit))
summary(y1model)

# Exposure-mediator model:

m1model <- glm(m1 ~ a1 + c1, family=gaussian(link=identity))
summary(m1model)

# Exposure model:

a1model <- glm(a1 ~ c1, family=gaussian(link=identity))
summary(a1model)

# Study 2 (mediator-outcome, mediator randomized):

c2 <- rnorm(nobs)
a2 <- rnorm(nobs, mean=alpha0 + alpha1 * c2, sd=1)
m2 <- rnorm(nobs, mean=beta0 + 0.0 * a2 + 0.0 * c2, sd=1)
py2 <- expit(theta0 + theta1 * a2 + theta2 * m2 + theta3 * a2 * m2 + theta4 * c2)

```

```

y2 <- rbinom(nobs, 1, py2)
table(y2)/nobs

# Full model (not available):

model2 <- glm(y2 ~ a2 + m2 + a2 * m2 + c2, family=binomial(link=logit))
summary(model2)

# Marginal mediator-outcome model:

y2model <- glm(y2 ~ m2 + c2, family=binomial(link=logit))
summary(y2model)

# Calculate theta1 at given level of theta2:

gettheta1 <- function(theta2val, theta3hat, theta1dag, beta0hat, beta1hat) {
  ma <- function(x, aval) {
    checkfinite <- exp((theta2val + theta3hat * aval) * x)
    lp <- beta0hat + beta1hat * aval
    return(ifelse(is.finite(checkfinite), checkfinite, .Machine$double.xmax) *
      dnorm(x, mean=beta0hat + beta1hat * aval, sd=msd))
  }
  ma0 <- integrate(ma, lower=-Inf, upper=Inf, aval=0)$value
  ma1 <- integrate(ma, lower=-Inf, upper=Inf, aval=1)$value
  return(theta1dag + log(ma0/ma1))
}

# Function to calculate difference between theta2star and its estimate (the root of
this is the estimate of theta2):

getdiff <- function(theta2val, theta3hat, theta1dag, theta2star, beta0hat, beta1hat,
alpha0hat, asd) {
  am <- function(x, mval) {
    checkfinite <- exp((theta1hat + theta3hat * mval) * x)
    return(ifelse(is.finite(checkfinite), checkfinite, .Machine$double.xmax) *
      dnorm(x, mean=alpha0hat, sd=asd))
  }
  results <- rep(NA, length(theta2val))
  for (i in 1:length(theta2val)) {
    theta1hat <- gettheta1(theta2val[i], theta3hat, theta1dag, beta0hat, beta1hat)
    am0 <- integrate(am, lower=-Inf, upper=Inf, mval=0)$value
    am1 <- integrate(am, lower=-Inf, upper=Inf, mval=1)$value
    results[i] <- theta2val[i] + log(am1/am0) - theta2star
  }
  return(results)
}

# Choose a value for the exposure-mediator interaction (true value used for
demonstration):

theta3val <- 0.25

# Inputs from the fitted models:

```

```

alpha0hat <- coef(a1model)[1]
asd <- sigma(a1model)
beta0hat <- coef(m1model)[1]
beta1hat <- coef(m1model)[2]
msd <- sigma(m1model)
theta1dag=coef(y1model)[2]
theta2star=coef(y2model)[2]

# Plot the function:

grid <- seq(-1, 1, by=0.1)
y <- getdiff(grid, theta3hat=theta3val, theta1dag=theta1dag, theta2star=theta2star,
  beta0hat=beta0hat, beta1hat=beta1hat, alpha0hat=alpha0hat, asd=asd)
plot(grid, y, type='l', xlab=expression(theta[2]), ylab='Difference', ylim=c(-1,1))
abline(h=0.0, lty='dashed')

# Solve for theta2 (main effect of mediator on outcome):

theta2hat <- uniroot(getdiff, interval=c(-1,3), theta3hat=theta3val,
  theta1dag=theta1dag, theta2star=theta2star,
  beta0hat=beta0hat, beta1hat=beta1hat, alpha0hat=alpha0hat,
  asd=asd)$root
abline(v=theta2hat, lty='dashed')
cat('\n', 'Estimated theta2=', round(theta2hat, 3), ' with theta3 fixed at ',
  theta3val, '\n\n', sep='')

# Get the corresponding theta1 (main effect of exposure on outcome):

theta1hat <- gettheta1(theta2hat, theta3val, theta1dag, beta0hat, beta1hat)
cat('\n', 'Estimated theta1=', round(theta1hat, 3), ' with theta3 fixed at ',
  theta3val, '\n\n', sep='')

# Natural direct and indirect effects for index versus reference level of exposure:
aind <- 1
aref <- 0

# Decomposition formulas for NIE and NDE estimates:

NIE <- (theta2hat*beta1hat + theta3val*beta1hat * aind) * (aind - aref)
NIE
NIE.OR <- exp(NIE); NIE.OR

NDE <- theta1hat + theta3val * (beta0hat + beta1hat * aref + theta2hat * msd^2) *
  (aind - aref) + 0.5 * theta3val^2 * msd^2 * (aind^2 - aref^2)
NDE
NDE.OR <- exp(NDE); NDE.OR

TE <- NIE + NDE
TE
TE.OR <- exp(TE); TE.OR

```