# Interventional effects for mediation analysis with multiple mediators: Supplementary Materials

March 22, 2016

## eAppendix A: Other scales

All results in the article extend immediately to other scales. Let $Y$be a dichotomous outcome coded 0 or 1. Then the effect (5) can be written as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}$$

on the relative risk scale, or as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}$$

on the odds ratio scale. The effect (6) can be written as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]}$$

on the relative risk scale, or as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{am_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a^*}=m_2|c)\right]}$$

on the odds ratio scale. The effect (7) can likewise be computed. Finally, the effect (8) can be written as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1, M_{2a}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}$$

on the relative risk scale, and as

$$\frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1, M_{2a}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{am_1m_2}|c\right) P(M_{1a}=m_1|c)P(M_{2a}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{am_1m_2}|c\right) P(M_{1a}=m_1, M_{2a}=m_2|c)\right]} \times \frac{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1, M_{2a^*}=m_2|c)\right]}{E\left[\sum_{m_1}\sum_{m_2} E\left(1-Y_{a^*m_1m_2}|c\right) P(M_{1a^*}=m_1|c)P(M_{2a^*}=m_2|c)\right]}$$

on the odds ratio scale. Each of the components of these effects can be calculated using the Monte-Carlo approach proposed in the main text of the article.

# eAppendix B: More than two mediators

With more than two mediators $M_1, ..., M_t$, the effect of exposure on outcome can be decomposed into many different path-specific effects. We choose not to infer all of these effects for the following two reasons. First, the scientific interest typically lies in knowing the effects that are mediated through each of the mediators, but rarely lies in all path-specific effects ways. Second, strong untestable assumptions are required to be able to infer all path-specific effects, such as assumptions about the direction of the causal effects between the various mediators, and about the absence of unmeasured common causes of all mediators. In this Appendix, we will therefore concentrate on the following pathways. We define the average interventional direct effect of exposure on outcome that is not via any of the mediators as:

$$E\left[\sum_{m_1}...\sum_{m_t}\left\{E\left(Y_{am_1...m_t}|c\right)-E\left(Y_{a^*m_1...m_t}|c\right)\right\}P(M_{1a^*}=m_1,...,M_{ta^*}=m_t|c)\right].$$

This expresses the effect of exposure on outcome when fixing the joint distribution of all mediators. It corresponds to the effect $A \rightarrow Y$ in the causal diagram of Figure 1 below.

For each mediator $M_s, s = 1, ..., t$, we further define the average interventional indirect effect via $M_s$ (but not its descendants) as

$$E\left[\sum_{m_1}...\sum_{m_t} E\left(Y_{am_1...m_t}|c\right)\left\{P(M_{sa}=m_s|c)-P(M_{sa^*}=m_s|c)\right\}\right.$$
$$\left.\times P(M_{1a}=m_1,...,M_{s-1,a}=m_{s-1}|c)P(M_{s+1,a^*}=m_{s+1},...,M_{ta^*}=m_t|c)\right]. \quad (1)$$

For $s = 1$, this corresponds to the effect $A \to M_1 \to Y$ in the causal diagram of Figure 1 below; for $s = 2$, this captures the combined effect along the pathways $A \to M_2 \to Y$ and $A \to M_1 \to M_2 \to Y$; for $s = 3$, it captures the combined effect along the pathways $A \to M_3 \to Y$, $A \to M_2 \to M_3 \to Y$, $A \to M_1 \to M_3 \to Y$ and $A \to M_1 \to M_2 \to M_3 \to Y$. Finally, it is easily seen that the difference between the total effect and the sum of the average interventional direct effect and the average interventional indirect effect via each of the mediators, captures an indirect effect of the exposure on the dependence between the mediators. Further work is needed to understand if the latter effect can be further decomposed into effects mediated via the dependence between specific subsets of mediators.

# eAppendix C: More details on the data analysis

## Data

In this Section, we give more detailed information on the NYCRIS data. Our analyses are based on all 29,580 women diagnosed with malignant, invasive breast cancer from 2000 to 2006 (inclusive) in NYCRIS who have information on cancer stage at diagnosis recorded; a further 2,589 women are excluded since this information is missing. For simplicity, we consider a binary SES exposure ($A$) which is whether or not the woman resides (at diagnosis) in an area (Lower Super Output Area) classified as belonging to the two most affluent quintiles of the national income distribution as defined by the income domain of the Indices of Multiple Deprivation (IMD) 2001. Since we have no direct information on screening, our first mediator ($M_1$) is a vector comprising age at diagnosis and cancer stage at diagnosis, classified as early (TNM stage 1 or 2) or advanced (TNM stage 3 or 4), considered jointly. Age and stage at diagnosis are strongly associated, likely due to the influence of screening and (latent) age at onset. Information on surgical treatment, obtained from a routinely collected national hospital dataset (Hospital Episode Statistics or HES), allows us to classify women either as having 'major surgery' (axillary dissection or other axillary nodal procedures, breast conserving surgery, mastectomy, and plastic surgery) or 'minor or no surgery' (other surgical procedures and none). This is our second mediator, $M_2$. The considered outcome ($Y$) is one-year survival from the date of diagnosis.

Calendar year at diagnosis and region (Yorkshire and The Humber, North East or North West) are considered as baseline confounders ($C$). As regards the causal structure of the mediators, we know that $M_1$ precedes $M_2$ and yet we can't rule out that they share unmeasured common causes, thus a combination of Figure 1 and Figure 3 of the main paper might apply. A possible causal diagram for the NYCRIS data is shown in Figure 4 of the main paper.

## Sequential mediation analysis

We begin by performing the sequential mediation analysis described at the beginning of Section 3.1. First we note that with $C = \{\text{Region, Year at diagnosis}\}$ in lieu of $C$, assumptions (i')-(iii') hold if Figure 4 represents the underlying causal diagram (with $M_1$ in lieu of $M_1$). If we additionally assume (iv'), then we can identify the natural direct effect not mediated through either $M_1$ or $M_2$ or both using (3) and the corresponding natural indirect effect through either $M_1$ or $M_2$ using (4). To estimate (3) and (4) using the Monte-Carlo approach of Section 3.3, we need to fit a series of associational models:

Model 1: We fit a logistic regression model to one-year survival ($Y$) conditional on SES ($A$), Stage and Age at diagnosis ($M_1$), Treatment ($M_2$), and Region and Year of diagnosis ($C$) with all interactions between $A$, $M_1$ and $M_2$ included.

Model 2: We also fit a logistic regression model to Treatment ($M_2$) conditional on SES ($A$), Stage and Age at diagnosis ($M_1$), and Region and Year of diagnosis ($C$) with all interactions between $A$ and $M_1$ included.

Model 3: We also fit a logistic regression model to Stage at diagnosis (one component of $M_1$) conditional on SES ($A$), Age at diagnosis (the other component of $M_1$), and Region and Year of diagnosis ($C$) with the interaction between SES and Age at diagnosis included.

Model 4: Finally, we fit a linear regression model to Age at diagnosis conditional on SES and Region and Year of diagnosis.

Note that this particular mediation analysis (with $M_1$ and $M_2$ considered as joint mediators) does not require any assumptions about the causal structure of the mediators; however, our associational models need to allow for correlation between them, and this is why we include Age in the model for Stage and Age and Stage in the model for Treatment. Also note that due to the very large sample size, there is little benefit in terms of precision (and a potential danger in terms of bias) in trying to find more parsimonious associational models than the above. Finally note that when using these results in the Monte-Carlo simulations to estimate (3), we will use not only the fitted value of the conditional expectation of age at diagnosis given SES and the confounders, but also the assumption that the errors from this model follow a normal distribution.

Tables 1–4 below give the full results of the individual regression models fitted to $M_1$, $M_2$ and $Y$. We use these results as described in Section 3.3 to estimate (3) and (4). Under assumptions (i)–(iv) with $M_1$ in lieu of $M$, we can additionally perform a mediation analysis with $M_1$ as the only mediator. Note that this involves assuming that $U$ in Figure 4 does not exist. For this mediation analysis, models 3 and 4 above are used again, together with:

Model 1': A logistic regression model for one-year survival ($Y$) conditional on SES ($A$), Stage and Age at diagnosis ($M_1$), and Region and Year of diagnosis ($C$) with all interactions between $A$ and $M_1$ included.

Models 1 and 1' are likely incompatible. We do not consider this to be of grave additional concern in practice, over and above the already substantial concern over parametric model misspecification in general.

We then use a Monte-Carlo approach to estimate the right-hand side of (1) in the main text with $L$ empty and $M_1$ in lieu of $M$, which, under assumptions (i)–(iv) is the natural direct effect not through $M_1$. By subtracting from this the estimate of the natural direct effect not through either or both of the mediators, we obtain our sequential mediation analysis estimate of the natural indirect effect through $M_2$ alone.

## Multiple mediator analysis based on interventional effects

We now perform the multiple mediator analysis described in Section 3.2, again using Monte-Carlo simulation as described at the end of Section 3.3. Details are given in the eAppendix. We make assumptions (i')–(iii'). In addition to models 1–4 above, we also now use:

Model 2': A logistic regression model for treatment ($M_2$) conditional on SES ($A$) and Region and Year of diagnosis ($C$).

The reason for specifying this model – which may be incompatible with model 2 – is that (6)–(8) all involve the distribution of $M_{2a}$ given $C$, which can be substituted by the distribution of $M_2$ given $A$ and $C$ under assumption (iii). Displays (5), on the other hand, involves model 2, and (8) involves both. The results are given in Table 2.

## Limitations

Our analyses are included mainly for illustration, to show how the proposed method can be applied in a realistic setting, and to show that even the most complicated effect, namely the mediated dependence (8), can have a meaningful interpretation when considered in an applied context. In order to focus on these interpretational issues, we made several simplifications that could be relaxed in future analyses of these data to gain a deeper and more reliable understanding of the reasons underlying socio-economic discrepancies in

breast cancer survival. Dichotomising both mediators has likely led to diluting the indirect effects and inflating the direct effect. In addition, dichotomising SES, the exposure, may have led to missing some more subtle effects across the income distribution. Focussing only on one-year survival may also mean that a different picture relating to longer term survival has been missed. In future work, we plan to relax all these simplifications in a more comprehensive substantive analysis, which will also involve sensitivity analyses to assess the impact of dropping women with unobserved stage at diagnosis. Another important limitation is the likely presence of unmeasured confounding, particularly of $M_1$ and $Y$ by the latent age at disease onset, and of $M_2$ and $Y$ by comorbidities, not available to us in the NYCRIS data. Sensitivity analyses to detect the plausible impact of such unmeasured confounding, as well as the robustness to the choice of a normality assumption for the errors from the model for age at diagnosis should also be explored.

# eAppendix D: Stata code for the data analysis

```
gen xm1a=x*m1a
gen xm1b=x*m1b
gen xm2=x*m2
gen m1ab=m1a*m1b
gen m1am2=m1a*m2
gen m1bm2=m1b*m2
gen xm1ab=x*m1a*m1b
gen xm1am2=x*m1a*m2
gen xm1bm2=x*m1b*m2
gen m1abm2=m1a*m1b*m2
gen xm1abm2=x*m1a*m1b*m2

logit y x m1a m1b m2 xm1a xm1b xm2 m1ab m1am2 m1bm2 xm1ab xm1am2 xm1bm2 m1abm2 xm1abm2 i.c1 i.c2
logit m2 x m1a m1b xm1a xm1b m1ab xm1ab i.c1 i.c2
logit m1b x m1a xm1a i.c1 i.c2
reg m1a x i.c1 i.c2

cap program drop seqMC
cap program define seqMC, rclass
cap drop m1a_0-tce
qui set obs 6000000
qui replace c1=c1[_n-29580] if c1==.
qui replace c2=c2[_n-29580] if c2==.

qui reg m1a x i.c1 i.c2
qui gen m1a_0 = _b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)+_b[2002.c2]*(c2==2002)
+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004) +_b[2005.c2]*(c2==2005)+_b[2006.c2]*(c2==2006)
+e(rmse)*rnormal()
qui gen m1a_1 = _b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)+_b[2002.c2]*(c2==2002)
+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004) +_b[2005.c2]*(c2==2005)+_b[2006.c2]*(c2==2006)+_b[x]
+e(rmse)*rnormal()

qui logit m1b x m1a xm1a i.c1 i.c2
qui gen m1b_0 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)
+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0)))
qui gen m1b_1 = runiform()<1/(1+exp(-(_b[_cons] +_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)
+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1)))
```

```
qui logit m2 x m1a m1b xm1a xm1b m1ab xm1ab i.c1 i.c2
qui gen m2_0 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003) +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0+_b[m1b]*m1b_0+_b[m1ab]*m1a_0*m1b_0)))
qui gen m2_1 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003) +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1
+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1)))

qui gen m2_01 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003) +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_1+_b[m1b]*m1b_1+_b[m1ab]*m1a_1*m1b_1)))
qui gen m2_10 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003) +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0+(_b[m1ab]
+_b[xm1ab])*m1a_0*m1b_0)))

qui logit y x m1a m1b m2 xm1a xm1b xm2 m1ab m1am2 m1bm2 xm1ab xm1am2 xm1bm2 m1abm2 xm1abm2 i.c1 i.c2
*M1 and M2 as joint mediators
qui gen y_00 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0+_b[m1b]*m1b_0+_b[m2]*m2_0+_b[m1ab]*m1a_0*m1b_0
+_b[m1am2]*m1a_0*m2_0+_b[m1bm2]*m1b_0*m2_0+_b[m1abm2]*m1a_0*m1b_0*m2_0)))
qui gen y_10 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0+(_b[m2]
+_b[xm2])*m2_0+(_b[m1ab]+_b[xm1ab])*m1a_0*m1b_0+(_b[m1am2]+_b[xm1am2])*m1a_0*m2_0+
(_b[m1bm2]+_b[xm1bm2])*m1b_0*m2_0+(_b[m1abm2]+_b[xm1abm2])*m1a_0*m1b_0*m2_0)))
qui gen y_01 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_1+_b[m1b]*m1b_1+_b[m2]*m2_1+_b[m1ab]*m1a_1*m1b_1
+_b[m1am2]*m1a_1*m2_1+_b[m1bm2]*m1b_1*m2_1+_b[m1abm2]*m1a_1*m1b_1*m2_1)))
qui gen y_11 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003) +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1+(_b[m2]
+_b[xm2])*m2_1+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1+(_b[m1am2]+_b[xm1am2])*m1a_1*m2_1+
(_b[m1bm2]+_b[xm1bm2])*m1b_1*m2_1+(_b[m1abm2]+_b[xm1abm2])*m1a_1*m1b_1*m2_1)))

*M1 as the only mediator
qui gen y_10_b = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0
+(_b[m2]+_b[xm2])*m2_10+(_b[m1ab]+_b[xm1ab])*m1a_0*m1b_0+(_b[m1am2]+_b[xm1am2])*m1a_0*m2_10
+(_b[m1bm2]+_b[xm1bm2])*m1b_0*m2_10+(_b[m1abm2]+_b[xm1abm2])*m1a_0*m1b_0*m2_10)))

qui gen NDE_M1M2=y_10-y_00
qui gen NIE_M1M2=y_11-y_10
qui gen NDE_M1=y_10_b-y_00
qui gen NIE_M1=y_11-y_10_b
qui gen NIE_M2alone=y_10_b-y_10

qui logit y x i.c1 i.c2
qui gen y1=1/(1+exp(-(_b[_cons]+_b[x]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006))))
qui gen y0=1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)
+_b[2005.c2]*(c2==2005)+_b[2006.c2]*(c2==2006))))
qui gen tce=y1-y0
qui summ tce
return scalar tce=r(mean)
```

```
qui summ NDE_M1M2
return scalar NDE_M1M2=r(mean)
qui summ NIE_M1M2
return scalar NIE_M1M2=r(mean)
qui summ NDE_M1
return scalar NDE_M1=r(mean)
qui summ NIE_M1
return scalar NIE_M1=r(mean)
qui summ NIE_M2alone
return scalar NIE_M2alone=r(mean)
end


cap program drop MMintMC
cap program define MMintMC, rclass
cap drop m1a_0-tce
qui set obs 6000000
qui replace c1=c1[_n-29580] if c1==.
qui replace c2=c2[_n-29580] if c2==.

qui reg m1a x i.c1 i.c2
qui gen m1a_0 = _b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+e(rmse)*rnormal()
qui gen m1a_1 = _b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]
+e(rmse)*rnormal()

qui logit m1b x m1a xm1a i.c1 i.c2
qui gen m1b_0 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0)))
qui gen m1b_1 = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1)))

qui logit m2 x m1a m1b xm1a xm1b m1ab xm1ab i.c1 i.c2
qui gen m2_0_cond = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0+_b[m1b]*m1b_0+_b[m1ab]*m1a_0*m1b_0)))
qui gen m2_1_cond = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1
+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1)))

qui logit m2 x i.c1 i.c2
qui gen m2_0_marg = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006))))
qui gen m2_1_marg = runiform()<1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x])))

qui logit y x m1a m1b m2 xm1a xm1b xm2 m1ab m1am2 m1bm2 xm1ab xm1am2 xm1bm2 m1abm2 xm1abm2 i.c1 i.c2
qui gen y_000_7 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[m1a]*m1a_0+_b[m1b]*m1b_0+_b[m2]*m2_0_cond
+_b[m1ab]*m1a_0*m1b_0+_b[m1am2]*m1a_0*m2_0_cond+_b[m1bm2]*m1b_0*m2_0_cond
+_b[m1abm2]*m1a_0*m1b_0*m2_0_cond)))
qui gen y_100_7 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
```

```
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0
+(_b[m2]+_b[xm2])*m2_0_cond+(_b[m1ab]+_b[xm1ab])*m1a_0*m1b_0+(_b[m1am2]
+_b[xm1am2])*m1a_0*m2_0_cond+(_b[m1bm2]+_b[xm1bm2])*m1b_0*m2_0_cond+
(_b[m1abm2]+_b[xm1abm2])*m1a_0*m1b_0*m2_0_cond)))

qui gen y_110_8 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)/// +_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1
+(_b[m2]+_b[xm2])*m2_0_marg+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1
+(_b[m1am2]+_b[xm1am2])*m1a_1*m2_0_marg+(_b[m1bm2]
+_b[xm1bm2])*m1b_1*m2_0_marg+(_b[m1abm2]+_b[xm1abm2])*m1a_1*m1b_1*m2_0_marg)))
qui gen y_100_8 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0
+(_b[m2]+_b[xm2])*m2_0_marg+(_b[m1ab]+_b[xm1ab])*m1a_0*m1b_0
+(_b[m1am2]+_b[xm1am2])*m1a_0*m2_0_marg+(_b[m1bm2]+_b[xm1bm2])*m1b_0*m2_0_marg
+(_b[m1abm2]+_b[xm1abm2])*m1a_0*m1b_0*m2_0_marg)))

qui gen y_101_9 = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)_b[x]+(_b[m1a]+_b[xm1a])*m1a_0+(_b[m1b]+_b[xm1b])*m1b_0
+(_b[m2]+_b[xm2])*m2_1_marg+(_b[m1ab]+_b[xm1ab])*m1a_0*m1b_0
+(_b[m1am2]+_b[xm1am2])*m1a_0*m2_1_marg+(_b[m1bm2]+_b[xm1bm2])*m1b_0*m2_1_marg
+(_b[m1abm2]+_b[xm1abm2])*m1a_0*m1b_0*m2_1_marg)))

qui gen y_111_10cond = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1
+(_b[m2]+_b[xm2])*m2_1_cond+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1
+(_b[m1am2]+_b[xm1am2])*m1a_1*m2_1_cond+(_b[m1bm2]+_b[xm1bm2])*m1b_1*m2_1_cond
+(_b[m1abm2]+_b[xm1abm2])*m1a_1*m1b_1*m2_1_cond)))
qui gen y_111_10marg = 1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006)+_b[x]+(_b[m1a]+_b[xm1a])*m1a_1+(_b[m1b]+_b[xm1b])*m1b_1
+(_b[m2]+_b[xm2])*m2_1_marg+(_b[m1ab]+_b[xm1ab])*m1a_1*m1b_1
+(_b[m1am2]+_b[xm1am2])*m1a_1*m2_1_marg+(_b[m1bm2]+_b[xm1bm2])*m1b_1*m2_1_marg
+(_b[m1abm2]+_b[xm1abm2])*m1a_1*m1b_1*m2_1_marg)))

*display 7
qui gen d7=y_100_7-y_000_7

*display 8
qui gen d8=y_110_8-y_100_8

*display 9
qui gen d9=y_101_9-y_100_8

*display 10
qui gen d10=y_111_10cond-y_111_10marg-y_100_7+y_100_8

qui logit y x i.c1 i.c2
qui gen y1=1/(1+exp(-(_b[_cons]+_b[x]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006))))
qui gen y0=1/(1+exp(-(_b[_cons]+_b[2.c1]*(c1==2)+_b[3.c1]*(c1==3)+_b[2001.c2]*(c2==2001)
+_b[2002.c2]*(c2==2002)+_b[2003.c2]*(c2==2003)+_b[2004.c2]*(c2==2004)+_b[2005.c2]*(c2==2005)
+_b[2006.c2]*(c2==2006))))
qui gen tce=y1-y0
qui summ tce
return scalar tce=r(mean)
```

```
qui summ d7
return scalar d7=r(mean)
qui summ d8
return scalar d8=r(mean)
qui summ d9
return scalar d9=r(mean)
qui summ d10
return scalar d10=r(mean)
end

bootstrap r(tce) r(d7) r(d8) r(d9) r(d10), reps(1000): MMintMC
bootstrap r(tce) r(NDE_M1M2) r(NIE_M1M2) r(NDE_M1) r(NIE_M1) r(NIE_M2alone), reps(1000): seqMC
```

| | Estimate | SE | 95% CI | |
|---|---|---|---|---|
| | | | lower | upper |
| Baseline odds* | 23.74 | 3.51 | 17.77 | 31.72 |
| | | | | |
| Conditional odds ratios | | | | |
| SES | | | | |
| higher | 1.871 | 0.411 | 1.216 | 2.877 |
| Age at diagnosis (yrs)** | 0.931 | 0.005 | 0.920 | 0.942 |
| Stage | | | | |
| advanced | 0.060 | 0.009 | 0.045 | 0.079 |
| Treatment | | | | |
| major | 2.975 | 0.443 | 2.222 | 3.984 |
| | | | | |
| SES×Agediag | 0.988 | 0.010 | 0.968 | 1.009 |
| SES×Stage | 0.657 | 0.164 | 0.402 | 1.073 |
| SES×Treat | 0.954 | 0.257 | 0.563 | 1.617 |
| Agediag×Stage | 1.056 | 0.008 | 1.041 | 1.071 |
| Agediag×Treat | 1.008 | 0.009 | 0.992 | 1.025 |
| Stage×Treat | 2.140 | 0.409 | 1.472 | 3.111 |
| SES×Agediag×Stage | 1.003 | 0.013 | 0.978 | 1.028 |
| SES×Agediag×Treat | 1.002 | 0.016 | 0.971 | 1.033 |
| SES×Stage×Treat | 1.090 | 0.376 | 0.555 | 2.142 |
| Agediag×Stage×Treat | 0.978 | 0.011 | 0.956 | 1.001 |
| SES×Agediag×Stage×Treat | 1.012 | 0.022 | 0.970 | 1.056 |
| | | | | |
| Region | | | | |
| North-West | 0.774 | 0.115 | 0.579 | 1.035 |
| Yorks | 0.991 | 0.059 | 0.881 | 1.114 |
| | | | | |
| Year of diagnosis | | | | |
| 2001 | 0.830 | 0.088 | 0.674 | 1.022 |
| 2002 | 0.942 | 0.102 | 0.762 | 1.165 |
| 2003 | 1.019 | 0.109 | 0.827 | 1.256 |
| 2004 | 0.954 | 0.103 | 0.772 | 1.180 |
| 2005 | 1.006 | 0.108 | 0.815 | 1.243 |
| 2006 | 1.092 | 0.120 | 0.879 | 1.355 |

Table 1: Results of logistic regression of one-year survival ($Y$) on SES ($A$), Stage and Age at diagnosis ($M_1$), Treatment ($M_2$), and Region and Year of diagnosis ($C$) with all interactions between $A$, $M_1$ and $M_2$. One-yr survival is coded 1 for survival and 0 for death.

* estimated odds of survival for women diagnosed in the North East region in 2000, with low SES, age at diagnosis 62 years, early stage and minor or no surgery

** centred at the mean age at diagnosis (61.8 years)

| | Estimate | SE | 95% CI | |
|---|---|---|---|---|
| | | | lower | upper |
| Baseline odds* | 4.796 | 0.226 | 4.373 | 5.261 |
| | | | | |
| Conditional odds ratios | | | | |
| SES | | | | |
| higher | 0.725 | 0.026 | 0.677 | 0.777 |
| Age at diagnosis (yrs)** | 0.937 | 0.002 | 0.934 | 0.941 |
| Stage | | | | |
| advanced | 0.186 | 0.009 | 0.169 | 0.205 |
| | | | | |
| SES×Agediag | 1.033 | 0.003 | 1.027 | 1.038 |
| SES×Stage | 1.799 | 0.152 | 1.525 | 2.123 |
| Agediag×Stage | 1.014 | 0.004 | 1.007 | 1.021 |
| SES×Agediag×Stage | 0.974 | 0.006 | 0.962 | 0.985 |
| | | | | |
| Region | | | | |
| North-West | 1.806 | 0.155 | 1.526 | 2.138 |
| Yorks | 0.795 | 0.025 | 0.747 | 0.846 |
| | | | | |
| Year of diagnosis | | | | |
| 2001 | 1.089 | 0.061 | 0.976 | 1.214 |
| 2002 | 1.119 | 0.062 | 1.003 | 1.249 |
| 2003 | 1.248 | 0.069 | 1.120 | 1.390 |
| 2004 | 1.429 | 0.081 | 1.280 | 1.596 |
| 2005 | 1.411 | 0.079 | 1.265 | 1.575 |
| 2006 | 1.442 | 0.082 | 1.291 | 1.611 |

Table 2: Results of logistic regression of Treatment ($M_2$) on SES ($A$), Stage and Age at diagnosis ($M_1$), and Region and Year of diagnosis ($C$) with all interactions between $A$ and $M_1$. Treatment is coded 1 for major surgery and 0 for minor or no surgery.
* estimated odds of major surgery for women diagnosed in the North East region in 2000, with low SES, age at diagnosis 62 years and early stage.
** centred at the mean age at diagnosis (61.8 years)

| | Estimate | SE | 95% CI | |
|---|---|---|---|---|
| | | | lower | upper |
| Baseline odds* | 0.164 | 0.009 | 0.148 | 0.182 |
| | | | | |
| Conditional odds ratios | | | | |
| SES | | | | |
| higher | 0.757 | 0.029 | 0.702 | 0.816 |
| Age at diagnosis (yrs)** | 1.020 | 0.002 | 1.017 | 1.023 |
| | | | | |
| SES×Agediag | 1.002 | 0.003 | 0.996 | 1.007 |
| | | | | |
| Region | | | | |
| North-West | 0.655 | 0.066 | 0.538 | 0.797 |
| Yorks | 1.059 | 0.040 | 0.985 | 1.140 |
| | | | | |
| Year of diagnosis | | | | |
| 2001 | 0.917 | 0.062 | 0.804 | 1.047 |
| 2002 | 0.950 | 0.064 | 0.833 | 1.083 |
| 2003 | 0.951 | 0.062 | 0.837 | 1.082 |
| 2004 | 0.845 | 0.057 | 0.741 | 0.965 |
| 2005 | 0.872 | 0.058 | 0.765 | 0.994 |
| 2006 | 0.909 | 0.061 | 0.798 | 1.036 |

Table 3: Results of logistic regression of Stage at diagnosis (one component of $M_1$) on SES ($A$), Age at diagnosis (the other component of $M_1$), and Region and Year of diagnosis ($C$) including the interaction between SES and age at diagnosis. Stage at diagnosis is coded 1 for advanced and 0 for early.

* estimated odds of being diagnosed at an advanced stage for women diagnosed in the North East region in 2000, with low SES and aged 62 years at diagnosis.

** centred at the mean age at diagnosis (61.8 years)

| | Estimate | SE | 95% CI | |
|---|---|---|---|---|
| | | | lower | upper |
| Baseline mean (intercept)* | 61.36 | 0.247 | 60.88 | 61.85 |
| | | | | |
| Mean differences / slopes | | | | |
| SES | | | | |
| higher | –1.53 | 0.168 | –1.86 | –1.20 |
| | | | | |
| Region | | | | |
| North-West | –0.488 | 0.383 | –1.24 | 0.262 |
| Yorks | 0.442 | 0.170 | 0.109 | 0.775 |
| | | | | |
| Year of diagnosis | | | | |
| 2001 | 0.616 | 0.309 | 0.011 | 1.22 |
| 2002 | 0.620 | 0.309 | 0.014 | 1.22 |
| 2003 | 1.36 | 0.303 | 0.765 | 1.95 |
| 2004 | 0.737 | 0.303 | 0.142 | 1.33 |
| 2005 | 1.13 | 0.302 | 0.542 | 1.73 |
| 2006 | 0.958 | 0.305 | 0.360 | 1.56 |
| | | | | |
| Residual standard deviation | | | | |
| | 13.87 | | | |

Table 4: Results of linear regression of Age at diagnosis (one component of $M_1$, in years) on SES ($A$) and Region and Year of diagnosis ($C$).
* estimated mean age at diagnosis for women diagnosed in the North East region in 2000, with low SES.

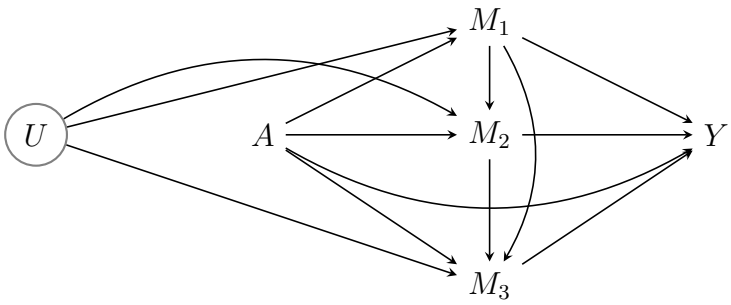


Figure 1: Causal diagram 5: multiple mediators.