eAppendix of "Bias formulas for sensitivity analysis for direct and indirect effects."

1. Bias Formulas for Sensitivity Analysis for an Unmeasured Confounder of the Exposure, Mediator and Outcome.

Here the setting is considered in which the unmeasured confounding variable U affects exposure A, mediator M and outcome Y so that U confounds both the exposure-outcome, mediator-outcome and exposure-mediator relationships as in Figure 2 in the text. Bias formulas for sensitivity analysis are given for controlled direct effects and for natural direct and indirect effects in this setting.

Theorem 3. Suppose that for all a and m, $Y_{am} \coprod A|\{C,U\}$ and $Y_{am} \coprod M|\{A,C,U\}$ then for any reference level u' of U we have that the differences between the average controlled direct effect $E[Y_{am} - Y_{a^*m}]$ and the biased estimator $\sum_c \{E[Y|a,m,c] - E[Y|a^*,m,c]\}P(c)$ is given by $Bias(CDE_{a,a^*}(m)) =$

$$\sum_c \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|c)\}P(c)$$
$$-\sum_c \sum_u \{E[Y|a^*,m,c,u] - E[Y|a^*,m,c,u']\}\{P(u|a^*,m,c) - P(u|c)\}P(c).$$

Under certain simplifying assumptions, the bias formula in Theorem 3 reduces to a simpler expression, given in Corollary 3, that is relatively straightforward to use in sensitivity analysis.

Corollary 3. Suppose that for all a and m, $Y_{am} \coprod A|\{C,U\}$ and $Y_{am} \coprod M|\{A,C,U\}$. Suppose further that U is binary, that $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0]$ is constant across strata of a,c so that $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0] = \gamma$ and that $P(U=1|a,m,c) - P(U=1|a^*,m,c)$ is constant across strata of c so that $P(u|a,m,c) - P(u|a^*,m,c) = \delta$ then

$$Bias(CDE_{a,a^*}(m)) = \delta\gamma.$$

Note that Corollary 3 does not require the assumption that $U \coprod A|C$ in Corollary 1. Theorem 4 and Corollary 4 give bias formulas for sensitivity analysis for natural direct and indirect effects allowing for U to affect A, M and Y. All of the no-unmeasured-confounding assumptions in Theorem 4 and Corollary 4 would be satisfied conditional on $\{C,U\}$, in the causal diagram in Figure 2 of the text.

Theorem 4. Suppose that for all a, a^*, and m, $Y_{am} \coprod A|\{C,U\}$, $Y_{am} \coprod M|\{A,C,U\}$, $M_a \coprod A|\{C,U\}$ and $Y_{am} \coprod M_{a^*}|\{C,U\}$ then for any reference level u' of U the bias formula for the natural direct effect is given by

$Bias(NDE_{a,a^*}(a^*)) =$

$$\sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c)P(c)$$
$$-\sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a^*,c) - P(u|c)\}P(m|a^*,c,u)P(c)$$

and the bias formula for the natural direct effect is given by $Bias(NIE_{a,a^*}(a)) =$

$$\sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,c,m) - \frac{P(u|c)}{P(u|a,c)}P(u|a,m,c)\}P(m|a,c)P(c)$$
$$-\sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \frac{P(u|c)}{P(u|a^*,c)}P(u|a^*,m,c)\}P(m|a^*,c)P(c)$$

As before, under certain simplifying assumptions, the bias formulas in Theorem 4 reduce to simpler expressions that are relatively straightforward to use in sensitivity analysis. Corollary 4 below gives these simple expressions and uses the same simplifying assumptions for natural direct and indirect effects.

Corollary 4. Suppose that for all a, a^*, and m, $Y_{am} \coprod A|\{C,U\}$, $Y_{am} \coprod M|\{A,C,U\}$, $M_a \coprod A|\{C,U\}$ and $Y_{am} \coprod M_{a^*}|\{C,U\}$. Suppose further that U is binary, that $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0]$ is constant across strata of a,m,c so that $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0] = \gamma$ and that $P(U=1|a,m,c) - P(U=1|a^*,m,c)$ is constant across strata of c so that $P(U=1|a,m,c) - P(U=1|a^*,m,c) = \delta_m$ then

$$\begin{aligned} Bias(NDE_{a,a^*}(a^*)) &= \gamma \sum_c \sum_m \delta_m P(m|a^*,c)P(c) \\ Bias(NIE_{a,a^*}(a^*)) &= \gamma \sum_c \sum_m P(U=1|a,m,c)\{P(m|a,c) - P(m|a^*,c)\}P(c) \end{aligned}$$

If $\delta_m = P(U=1|a,m,c) - P(U=1|a^*,m,c)$ is constant across strata of m taking value δ then $Bias(NDE_{a,a^*}(a^*)) = \delta\gamma$ and $Bias(NIE_{a,a^*}(a^*)) = -\delta\gamma - \gamma \sum_c \{P(U=1|a,c) - P(U=1|a^*,c)\}P(c)$.

Note that Corollary 4 gives the same bias formula for natural direct effects as Corollary 2 in the paper and thus the same approach as described in the paper for sensitivity analysis for natural direct effects may be used under the more general setting in which the unmeasured confounding variable U affects A, M and Y. However, for sensitivity analysis for natural indirect effects when U also affects A, the bias for natural indirect effects is not simply the negation of the bias for natural direct effects. This is because when U affects not just M and Y but also A, the total effect of A on Y is also confounded and thus $\sum_c \{E[Y|a,c] - E[Y|a^*,c]\}P(c)$ will be biased for the total effect.

2. General Bias Formulas for Controlled Direct Effect and Natural Direct and Indirect Effect Risk Ratios.

The following results generalize the simple bias formulas given in Appendix 1 concerning controlled direct effect and natural direct and indirect effect risk ratios.

Theorem 5. Suppose that for all a and m, $Y_{am} \coprod A|C$ and $Y_{am} \coprod M|\{A, C, U\}$ then for any reference level u' of U we have that

$$Bias(CDE^{RR}_{a,a^*|c}(m)) = \frac{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')} P(u|a,m,c)}{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')} P(u|a,c)} \Big/ \frac{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')} P(u|a^*,m,c)}{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')} P(u|a^*,c)}.$$

It follows from this, as noted in Appendix 1, that if U is binary, if $U \coprod A|C$, and if $\frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U=0)} = \gamma$ is constant across strata of a then

$$Bias(CDE^{RR}_{a,a^*|c}(m)) = \frac{1 + (\gamma - 1)P(U = 1|a, m, c)}{1 + (\gamma - 1)P(U = 1|a^*, m, c)}.$$

Theorem 6. If Figure 1 represents a causal directed graph then for all a, a^*, and m, $Y_{am} \coprod A|C$, $Y_{am} \coprod M|\{A, C, U\}$, $M_a \coprod A|C$ and $Y_{am} \coprod M_{a^*}|\{C, U\}$ and $U \coprod A|C$ and for any reference level u' of U and any reference level m' of M we have that

$$\begin{aligned} Bias(NDE^{RR}_{a,a^*|c}(a^*)) &= \frac{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']} P(u|a,m,c) \frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']} P(m|a^*,c)}{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']} P(u|a^*,m,c) \frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']} P(m|a^*,c)} \\ Bias(NIE^{RR}_{a,a^*|c}(a)) &= 1/Bias(NDE^{RR}_{a,a^*|c}(a^*)) \end{aligned}$$

It follows from this, as noted in Apendix 1, that if U is binary and if $\frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U,=0)} = \gamma$ is constant across strata of m then

$$\begin{aligned} Bias(NDE^{RR}_{a,a^*|c}(a^*)) &= \frac{\sum_m \{1 + (\gamma - 1)\pi_{a,m}\} v_m P(m|a^*, c)}{\sum_m \{1 + (\gamma - 1)\pi_{a^*,m}\} v_m P(m|a^*, c)} \\ Bias(NIE^{RR}_{a,a^*|c}(a)) &= 1/Bias(NDE^{RR}_{a,a^*|c}(a^*)). \end{aligned}$$

where $\pi_{a,m} = P(U = 1|a, m, c)$, $\pi_{a^*,m} = P(U = 1|a^*, m, c)$ and $v_m = \frac{E[Y|a,m,c,U=0]}{E[Y|a,m',c,U=0]}$. If $\pi_{a,m}$ and $\pi_{a^*,m}$ are constant across m so that $\pi_{a,m} = \pi_a$ and $\pi_{a^*,m} = \pi_{a^*}$ and if $v_m = 1$ for all m then

$$\begin{aligned} Bias(NDE^{RR}_{a,a^*|c}(a^*)) &= \frac{1 + (\gamma - 1)\pi_a}{1 + (\gamma - 1)\pi_{a^*}} \\ Bias(NIE^{RR}_{a,a^*|c}(a)) &= \frac{1 + (\gamma - 1)\pi_{a^*}}{1 + (\gamma - 1)\pi_a}. \end{aligned}$$

3. Bias Formulas for Sensitivity Analysis for Principal Strata Direct Effects.

Suppose that the exposure A and the mediator M are binary. The principal strata direct effects are defined by $PSDE(m) = E(Y_1 - Y_0|M_1 = M_0 = m)$ i.e. the principal strata direct effect for principal strataum $(M_1 = m, M_0 = m)$ is defined as what the effect of the exposure on the outcome would be amongst those individuals for whom the mediator level would be m irrespective of whether the exposure was $A = 1$ or $A = 0$.

If the effect of A on M is monotonic in the sense that $M_0 \leq M_1$ for all individuals in the population and if for all a, a^*, m, $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ as would hold if A were randomized and if assumption (2) in the text holds that $Y_{am} \coprod M|\{A, C\}$ then principal strata direct effects are identified as stated in the following Theorem which also gives bias formulas for the principal strata direct effect if there is an unmeasured confounding variable U.

Theorem 7. If $M_0 \leq M_1$ for all individuals in the population and if for all a, a^*, m we have $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ and $Y_{am} \coprod M|\{A, C\}$ then

$$E(Y_1 - Y_0|M_1 = M_0 = m, C = c) = E(Y|A = 1, M = m, C = c) - E(Y|A = 0, M = m, C = c).$$

If there is an unmeasured confounder U such that $\{Y_{am}, M_a, M_{a^*}\} \coprod A|\{C, U\}$ and $Y_{am} \coprod M|\{A, C, U\}$ then for any reference level u' we have that

$$\begin{aligned} E(Y_1 - Y_0|M_1 = M_0 = m, C = c) \quad = \quad & E(Y|A = 1, m, c) - E(Y|A = 0, m, c) \\ & - [\textstyle\sum_u \{E[Y|A = 1, m, c, u] - E[Y|A = 1, m, c, u']\}\{P(u|A = 1, m, c) - P(u|c)\} \\ & - \textstyle\sum_u \{E[Y|A = 0, m, c, u] - E[Y|A = 0, m, c, u']\}\{P(u|A = 0, m, c) - P(u|c)\}]. \end{aligned}$$

Moreover if U is binary with $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0]$ constant across strata of a, c so that $E[Y|a, m, c, U = 1] - E[Y|a, m, c, U = 0] = \gamma$ and $P(U = 1|A = 1, m, c) - P(U = 1|A = 0, m, c)$ constant across strata of c so that $P(u|A = 1, m, c) - P(u|A = 0, m, c) = \delta$ then

$$E(Y_1 - Y_0|M_1 = M_0 = m, C = c) = E(Y|A = 1, m, c) - E(Y|A = 0, m, c) - \delta\gamma.$$

As shown in the proof of Theorem 7 below if we wish to avoid making reference to counterfactuals of the form Y_{am}, the identification formula in Theorem 7 also holds if instead of assuming $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ and $Y_{am} \coprod M|\{A, C\}$ we assume $\{Y_a, M_{a^*}\} \coprod A|C$ and $Y_a \coprod A|\{M = a, C\}$. The bias formulas in Theorem 7 will still hold if instead of assuming $\{Y_{am}, M_a, M_{a^*}\} \coprod A|\{C, U\}$ and $Y_{am} \coprod M|\{A, C, U\}$ we assume $\{Y_a, M_{a^*}\} \coprod A|\{C, U\}$ and $Y_a \coprod A|\{M = a, C, U\}$.

4. Example using Bias Formulas for Direct Effect Risk Ratios.

We give an example employing the bias formula results for direct effect risk ratios illustrating a case in which the unmeasured mediator-outcome confounder may completely explain away the apparent direct effect. Empirical studies have found that when controlling for birth weight, M, for the group of infants with the lowest birth weight ($M = 0$), maternal smoking, A, is associated with a lower risk of infant mortality, Y, seemingly suggesting a protective effect for maternal smoking amongst infants weighing the least; this somewhat puzzling finding is commonly referred to as the "birth weight paradox"[25,45,46]. Hernández-Díaz et al.[25] point out that although analyses that document this association control for a number of maternal demographic factors, C, the analyses do not in general control for birth defects or malnutrition, U (i.e. other causes of low birth weight), which would serve as a confounder of the birth weight (mediator) - mortality (outcome) relationship. Estimates of the controlled direct effect are thus biased because control is not made for such mediator-outcome confounders. Essentially infants might be low birth weight either because of smoking or because of say a birth defect or malnutrition. If an infant is not low birth weight because of smoking it is more likely that the low birth weight is because of a birth defect or malnutrition or some other cause. Thus, if control is not made for these other causes of low birth weight and comparison is made between the groups with and without maternal smoking it looks as if the effect of smoking is protective for the infants of lowest birth weight; this is simply because for this group of low birth weight infants, control is not made for other causes of low birth weight and thus no smoking and low birth weight together is likely indicative of the presence of a birth defect or malnutrition. Using the sensitivity analysis techniques we can assess the degree of confounding required to completely explain away the birth weight paradox. Hernández-Díaz et al.[25] use 1991 US linked birth/infant-death data sets from the National Center for Health Statistics and they define the lowest birth weight category ($M = 0$) as birth weight less than 2,000g. They control for maternal age, gravidity, education, marital status, race/ethnicity, and prenatal care (denoted by C). They find that if a naive estimate is used of the controlled direct effect risk ratio, one obtains $\frac{P(Y=1|A=1,M=0,c)}{P(Y=1|A=0,M=0,c)} = 0.79$, suggesting a protective effect of smoking for the birth weight in stratum $M = 0$. Suppose, for sensitivity analysis, that we take U to be other causes of low birth weight then using the bias formulas for risk ratios presented in Appendix 1, we find that if $U = 1$ were to conditionally increases the risk of infant mortality three-and-a-half-fold so that $\frac{P(Y|a,m,c,U=1)}{P(Y|a,m,c,U=0)} = 3.5$ and if the prevalence of U for low-birth weight infants whose mothers smoke is 0.025 but the prevalence of U for low-birth weight infants whose mothers do not smoke is 0.14 (smoking is ruled out as an explanation of low-birth weight rendering other causes more likely) then this would indicate the bias produced by the unmeasured mediator-outcome confounder U, namely $\frac{1+(3.5-1)(0.03)}{1+(3.5-1)(0.14)} = 0.79$, is sufficient to completely explain the apparent protective effect. See Whitcomb et al.[46] for a related simulation-based analysis of the birth-weight paradox.

5. Proofs.

Proofs of Theorem 3 and Corollary 3.

The proof of Theorem 3 proceeds exactly as Theorem 1 but with both the expressions $P(u|a,c)$ and $P(u|a^*,c)$ replaced by $P(u|c)$ throughout the proof. The proof of Corollary 3 proceeds in a similar manner to that of Corollary 1. Note because in bias expressions given in Theorem 3, the terms $P(u|a,c)$ and $P(u|a^*,c)$ are both replaced by $P(u|c)$, we do not have to make the assumption that $U \coprod A|C$ in Corollary 3.

Proofs of Theorems 2 and 4.

Note that in the causal directed acyclic graph in Figure 1 in the paper, the conditions of Theorem 4 are satisfied. For the natural direct effect we have that, $Bias(NDE_{a,a^*}(a^*))$

$$
\begin{aligned}
&= \sum_c \sum_m \{E[Y|a,m,c] - E[Y|a^*,m,c]\}P(m|a^*,c)P(c) - E[Y_{aM_{a^*}} - Y_{a^*M_{a^*}}] \\
&= \sum_c \sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)P(c) - \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]P(u|a^*,m,c)P(m|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u E[Y|a,m,c,u]P(m|a^*,c,u)P(u|c)P(c) + \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)P(c) \\
&= \sum_c \sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)P(c) - \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]\tfrac{P(m|a^*,c,u)P(u|a^*,c)}{P(m|a^*,c)}P(m|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u E[Y|a,m,c,u]\tfrac{P(u|a^*,m,c)P(m|a^*,c)}{P(u|a^*,c)}P(u|c)P(c) + \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)P(c) \\
&= \sum_c \sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)P(c) - \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u E[Y|a,m,c,u]\tfrac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}P(m|a^*,c)P(c) + \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)P(c) \\
&= \sum_c \sum_m \sum_u E[Y|a,m,c,u]\{P(u|a,m,c) - \tfrac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u E[Y|a^*,m,c,u]\{P(u|a^*,c) - P(u|c)\}P(m|a^*,c,u)P(c) \\
&= \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \tfrac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a^*,c) - P(u|c)\}P(m|a^*,c,u)P(c).
\end{aligned}
$$

where the second equality follows because $Y_{am} \coprod A|\{C,U\}$, $Y_{am} \coprod M|\{A,C,U\}$, $M_a \coprod A|\{C,U\}$ and $Y_{am} \coprod M_{a^*}|\{C,U\}$ (see for example Pearl[3] or the eAppendix in VanderWeele[21]). This proves the part of Theorem 4 concerning natural direct effects. To prove the part of Theorem 2 concerning natural direct effects, note that in the causal directed acyclic graph in Figure 1 we have that $U \coprod A|C$ and thus $Bias(NDE_{a,a^*}(a^*))$

$$
\begin{aligned}
&= \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \frac{P(u|a^*,m,c)P(u|c)}{P(u|c)}\}P(m|a^*,c)P(c) \\
&\quad - \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|c) - P(u|c)\}P(m|a^*,c,u)P(c) \\
&= \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - P(u|a^*,m,c)\}P(m|a^*,c)P(c).
\end{aligned}
$$

For the natural indirect effects we have that, $Bias(NIE_{a,a^*}(a))$

$$
\begin{aligned}
&= \sum_c \sum_m E[Y|a,m,c]\{P(m|a,c) - P(m|a^*,c)\}P(c) - E[Y_{aM_a} - Y_{aM_{a^*}}] \\
&= \sum_c \sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)\{P(m|a,c) - P(m|a^*,c)\}P(c)
\end{aligned}
$$

$$
\begin{aligned}
&-\sum_c\sum_m\sum_u E[Y|a,m,c,u]P(m|a,c,u)P(u|c)P(c)+\sum_c\sum_m\sum_u E[Y|a,m,c,u]P(m|a^*,c,u)P(u|c)P(c)\\
&=\sum_c\sum_m\sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a,c)P(c)-\sum_c\sum_m\sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)\}P(c)\\
&-\sum_c\sum_m\sum_u E[Y|a,m,c,u]\tfrac{P(u|a,m,c)}{P(u|a,c)}P(m|a,c)P(u|c)P(c)+\sum_c\sum_m\sum_u E[Y|a,m,c,u]\tfrac{P(u|a^*,m,c)}{P(u|a^*,c)}P(m|a^*,c)P(u|c)P(c)\\
&=\sum_c\sum_m\sum_u E[Y|a,m,c,u]\{P(u|a,m,c)-\tfrac{P(u|c)}{P(u|a,c)}P(u|a,m,c)\}P(m|a,c)P(c)\\
&-\sum_c\sum_m\sum_u E[Y|a,m,c,u]\{P(u|a,m,c)-\tfrac{P(u|c)}{P(u|a^*,c)}P(u|a^*,m,c)\}P(m|a^*,c)P(c)\\
&=\sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a,m,c)-\tfrac{P(u|c)}{P(u|a,c)}P(u|a,m,c)\}P(m|a,c)P(c)\\
&-\sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a,m,c)-\tfrac{P(u|c)}{P(u|a^*,c)}P(u|a^*,m,c)\}P(m|a^*,c)P(c)
\end{aligned}
$$

where the second equality follows because $Y_{am} \coprod A|\{C,U\}$, $Y_{am} \coprod M|\{A,C,U\}$, $M_a \coprod A|\{C,U\}$ and $Y_{am} \coprod M_{a^*}|\{C,U\}$ (see for example Pearl[3] or the eAppendix in VanderWeele[21]). This proves the part of Theorem 4 concerning natural indirect effects. To prove the part of Theorem 2 concerning natural indirect effects, note that if $U \coprod A|C$ and thus $Bias(NIE_{a,a^*}(a))$

$$
\begin{aligned}
&= \sum_c\sum_m\sum_u E[Y|a,m,c,u]\{P(u|a,m,c)-\frac{P(u|c)}{P(u|c)}P(u|a,m,c)\}P(m|a,c)P(c)\\
&\quad -\sum_c\sum_m\sum_u E[Y|a,m,c,u]\{P(u|a,m,c)-\frac{P(u|c)}{P(u|c)}P(u|a^*,m,c)\}P(m|a^*,c)P(c)\\
&= -\sum_c\sum_m\sum_u E[Y|a,m,c,u]\{P(u|a,m,c)-P(u|a^*,m,c)\}P(m|a^*,c)P(c)\\
&= -\sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a,m,c)-P(u|a^*,m,c)\}P(m|a^*,c)P(c).
\end{aligned}
$$

This completes the proof.

Proof of Corollary 4.

From Theorem 4, we have that $Bias(NDE_{a,a^*}(a^*))$

$$
\begin{aligned}
&= \sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a,m,c)-\frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c)P(c)\\
&\quad -\sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a^*,c)-P(u|c)\}P(m|a^*,c,u)P(c)\\
&= \sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a,m,c)-\frac{P(u|a^*,m,c)P(u|c)}{P(u|a^*,c)}\}P(m|a^*,c)P(c)\\
&\quad -\sum_c\sum_m\sum_u \{E[Y|a,m,c,u]-E[Y|a,m,c,u']\}\{P(u|a^*,c)-P(u|c)\}\frac{P(u|a^*,m,c)}{P(u|a^*,c)}P(m|a^*,c)P(c)\\
&= \sum_c\sum_m \gamma\{P(U=1|a,m,c)-\frac{P(U=1|a^*,m,c)P(U=1|c)}{P(U=1|a^*,c)}\}P(m|a^*,c)P(c)\\
&\quad -\sum_c\sum_m \gamma\{P(U=1|a^*,c)-P(U=1|c)\}\frac{P(U=1|a^*,m,c)}{P(U=1|a^*,c)}P(m|a^*,c)P(c)\\
&= \gamma\sum_c\sum_m \{P(U=1|a,m,c)-P(U=1|a^*,m,c)\}P(m|a^*,c)P(c)\\
&= \gamma\sum_c\sum_m \delta_m P(m|a^*,c)P(c).
\end{aligned}
$$

If for all m, $\delta_m = P(U=1|a,m,c) - P(U=1|a^*,m,c)$ takes value δ then $Bias(NDE_{a,a^*}(a^*)) = \gamma \sum_c \sum_m \delta P(m|a^*,c)P(c) = \delta\gamma$. This gives the result for natural direct effects. From Theorem 4 we have for natural indirect effects that, $Bias(NIE_{a,a^*}(a))$

$$
\begin{aligned}
= \quad & \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \frac{P(u|c)}{P(u|a,c)} P(u|a,m,c)\}P(m|a,c)P(c) \\
& - \sum_c \sum_m \sum_u \{E[Y|a,m,c,u] - E[Y|a,m,c,u']\}\{P(u|a,m,c) - \frac{P(u|c)}{P(u|a^*,c)} P(u|a^*,m,c)\}P(m|a^*,c)P(c) \\
= \quad & \gamma \sum_c \sum_m \{P(U=1|a,m,c) - \frac{P(U=1|c)}{P(U=1|a,c)} P(U=1|a,m,c)\}P(m|a,c)P(c) \\
& - \gamma \sum_c \sum_m \{P(U=1|a,m,c) - \frac{P(U=1|c)}{P(U=1|a^*,c)} P(U=1|a^*,m,c)\}P(m|a^*,c)P(c) \\
= \quad & \gamma \sum_c \sum_m P(U=1|a,m,c)P(m|a,c)P(c) - \gamma \sum_c \sum_m P(U=1|c)P(m|a,c,U=1)P(c) \\
& - \gamma \sum_c \sum_m P(U=1|a,m,c)P(m|a^*,c)P(c) - \gamma \sum_c \sum_m P(U=1|c)P(m|a^*,c,U=1)P(c) \\
= \quad & \gamma \sum_c \sum_m P(U=1|a,m,c)\{P(m|a,c) - P(m|a^*,c)\}P(c).
\end{aligned}
$$

If for all m, $\delta_m = P(U=1|a,m,c) - P(U=1|a^*,m,c)$ takes value δ then we have that $Bias(NIE_{a,a^*}(a))$

$$
\begin{aligned}
= \quad & \gamma \sum_c \sum_m P(U=1|a,m,c)\{P(m|a,c) - P(m|a^*,c)\}P(c) \\
= \quad & \gamma \sum_c \sum_m \{P(U=1|a,m,c) - \delta\}\{P(m|a,c) - P(m|a^*,c)\}P(c) \\
= \quad & \gamma \sum_c \sum_m \{P(U=1|a,m,c) - \delta\}P(m|a,c)P(c) \\
& - \gamma \sum_c \sum_m \{P(U=1|a^*,m,c)\}P(m|a^*,c)P(c) \\
= \quad & \gamma \sum_c P(U=1|a,c)P(c) - \delta\gamma \\
& - \gamma \sum_c P(U=1|a^*,c)P(c) \\
= \quad & -\delta\gamma - \gamma \sum_c \{P(U=1|a,c) - P(U=1|a^*,c)\}P(c)
\end{aligned}
$$

This completes the proof.

Proof of Theorem 5 and the Simple Formula in Appendix 1.

We have for any reference level u' that

$$\begin{aligned}
Bias(CDE^{RR}_{a,a^*}(m)) &= \frac{E(Y|a,m,c)/E(Y|a^*,m,c)}{E(Y_{am}|c)/E(Y_{a^*m}|c)} \\
&= \frac{\sum_u E(Y|a,m,c,u)P(u|a,m,c)}{\sum_u E(Y_{am}|c,u)P(u|a,c)} / \frac{\sum_u E(Y|a^*,m,c,u)P(u|a^*,m,c)}{\sum_u E(Y_{a^*m}|c,u)P(u|a^*,c)} \\
&= \frac{\sum_u E(Y|a,m,c,u)P(u|a,m,c)}{\sum_u E(Y|a,m,c,u)P(u|a,c)} / \frac{\sum_u E(Y|a^*,m,c,u)P(u|a^*,m,c)}{\sum_u E(Y|a^*,m,c,u)P(u|a^*,c)} \\
&= \frac{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')}P(u|a,m,c)}{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')}P(u|a,c)} / \frac{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')}P(u|a^*,m,c)}{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')}P(u|a^*,c)}.
\end{aligned}$$

This proves the general formula in Theorem 5. To derive the simple bias formula for controlled direct effect risk ratios given in Appendix 1, if $U \coprod A|C$ and we let $u' = 0$ then we have

$$\begin{aligned}
Bias(CDE^{RR}_{a,a^*}(m)) &= \frac{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')}P(u|a,m,c)}{\sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')}P(u|c)} / \frac{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')}P(u|a^*,m,c)}{\sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')}P(u|c)} \\
&= \sum_u \frac{E(Y|a,m,c,u)}{E(Y|a,m,c,u')}P(u|a,m,c) / \sum_u \frac{E(Y|a^*,m,c,u)}{E(Y|a^*,m,c,u')}P(u|a^*,m,c) \\
&= \frac{\gamma P(U=1|a,m,c) + P(U=0|a,m,c)}{\gamma P(U=1|a^*,m,c) + P(U=0|a^*,m,c)} \\
&= \frac{1+(\gamma-1)P(U=1|a,m,c)}{1+(\gamma-1)P(U=1|a^*,m,c)}.
\end{aligned}$$

This completes the proof.

Proof of Theorem 6 and the Simple Formulas in Appendix 1.

We have for any reference level u' that $Bias(NDE^{RR}_{a,a^*|c}(a^*))$

$$\begin{aligned}
&= \frac{\sum_m P(Y|a,m,c)P(m|a^*,c) / \sum_m P(Y|a^*,m,c)P(m|a^*,c)}{P(Y_{aM_{a^*}}|c)/P(Y_{a^*M_{a^*}}|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) / \sum_m \sum_u E[Y|a^*,m,c,u]P(u|a^*,m,c)P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]P(m|a^*,c,u)P(u|c) / \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) / \sum_m \sum_u E[Y|a^*,m,c,u]\frac{P(m|a^*,c,u)P(u|a^*,c)}{P(m|a^*,c)}P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]\frac{P(u|a^*,m,c)P(m|a^*,c)}{P(u|a^*,c)}P(u|c) / \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c) / \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)}{\sum_m \sum_u E[Y|a,m,c,u]P(u|a^*,m,c)P(m|a^*,c) / \sum_m \sum_u E[Y|a^*,m,c,u]P(m|a^*,c,u)P(u|c)} \\
&= \frac{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']}P(u|a,m,c)\frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']}P(m|a^*,c)}{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']}P(u|a^*,m,c)\frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']}P(m|a^*,c)}
\end{aligned}$$

where the second equality follows because $Y_{am} \coprod A|\{C,U\}$, $Y_{am} \coprod M|\{A,C,U\}$, $M_a \coprod A|\{C,U\}$ and $Y_{am} \coprod M_{a^*}|\{C,U\}$ and the fourth equality follows because $U \coprod A|C$. For the natural indirect effect we have that $Bias(NIE^{RR}_{a,a^*|c}(a))$

$$
\begin{aligned}
&= \frac{\sum_m P(Y|a,m,c)P(m|a,c)/\sum_m P(Y|a,m,c)P(m|a^*,c)}{P(Y_{aM_a}|c)/P(Y_{aM_{a^*}}|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a,c)/\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]P(m|a,c,u)P(u|c)/\sum_m \sum_u E[Y|a,m,c,u]P(m|a^*,c,u)P(u|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a,c)/\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]\frac{P(u|a,m,c)}{P(u|a,c)}P(m|a,c)P(u|c)/\sum_m \sum_u E[Y|a,m,c,u]\frac{P(u|a^*,m,c)}{P(u|a^*,c)}P(m|a^*,c)P(u|c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a,c)/\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a,c)/\sum_m \sum_u E[Y|a,m,c,u]P(u|a^*,m,c)P(m|a^*,c)} \\
&= \frac{\sum_m \sum_u E[Y|a,m,c,u]P(u|a^*,m,c)P(m|a^*,c)}{\sum_m \sum_u E[Y|a,m,c,u]P(u|a,m,c)P(m|a^*,c)} \\
&= \frac{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']}P(u|a^*,m,c)\frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']}P(m|a^*,c)}{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,u']}P(u|a,m,c)\frac{E[Y|a,m,c,u']}{E[Y|a,m',c,u']}P(m|a^*,c)} \\
&= 1/Bias(NDE^{RR}_{a,a^*|c}(a^*))
\end{aligned}
$$

This proves the general formulas in Theorem 6. To derive the simple bias formula for the natural direct effect risk ratio given in Appendix 1, let $u' = 0$ then we have

$$
\begin{aligned}
Bias(NDE^{RR}_{a,a^*|c}(a^*)) &= \frac{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,U=0]}P(u|a,m,c)v_m P(m|a^*,c)}{\sum_m \sum_u \frac{E[Y|a,m,c,u]}{E[Y|a,m,c,U=0]}P(u|a^*,m,c)v_m P(m|a^*,c)} \\
&= \frac{\sum_m \{\gamma P(U=1|a,m,c) + P(U=0|a,m,c)\}v_m P(m|a^*,c)}{\sum_m \{\gamma P(U=1|a^*,m,c) + P(U=0|a,m,c)\}v_m P(m|a^*,c)} \\
&= \frac{\sum_m \{1 + (\gamma-1)P(U=1|a,m,c)\}v_m P(m|a^*,c)}{\sum_m \{1 + (\gamma-1)P(U=1|a^*,m,c)\}v_m P(m|a^*,c)} \\
&= \frac{\sum_m \{1 + (\gamma-1)\pi_{a,m}\}v_m P(m|a^*,c)}{\sum_m \{1 + (\gamma-1)\pi_{a^*,m}\}v_m P(m|a^*,c)}.
\end{aligned}
$$

Furthermore if $\pi_{a,m}$ and $\pi_{a^*,m}$ are constant across m and if $v_m = 1$ for all m then this reduces to

$$= \frac{1 + (\gamma-1)\pi_a}{1 + (\gamma-1)\pi_{a^*}}.$$

This completes the proof.

Proof of Theorem 7.

If If $M_0 \leq M_1$ for all individuals in the population and $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ and $Y_{am} \coprod M|\{A, C\}$ then

$$E(Y|A=1,M=0,c) = E(Y_{10}|A=1,M=0,c) = E(Y_{10}|A=1,M_1=0,c) = E(Y_{10}|M_1=0,c) =$$
$$= E(Y_{10}|M_1=M_0=0,c) = E(Y_{1M_1}|M_1=M_0=0,c) = E(Y_1|M_1=M_0=0,c)$$

where the first and second equalities follow by consistency, the third by the assumption $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$, the fourth by monotonicity and the final one by composition. We also have that

$$E(Y|A=0,M=0,c) = E(Y_{00}|A=0,M=0,c) = E(Y_{00}|A=0,c) = E(Y_{00}|A=1,c) = E(Y_{00}|A=1,M=0,c)$$
$$= E(Y_{00}|A=1,M_1=0,c) = E(Y_{00}|A=1,M_1=M_0=0,c) = E(Y_{00}|M_1=M_0=0,c)$$
$$= E(Y_{0M_0}|M_1=M_0=0,c) = E(Y_0|M_1=M_0=0,c)$$

where the first and the fifth equalities follow by consistency, the second and fourth by the assumption $Y_{am} \coprod M|\{A, C\}$, the third and the seventh by the assumption $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$, the sixth by monotonicity and the final one by composition. Similarly,

$$E(Y|A=1,M=1,c) = E(Y_{11}|A=1,M=1,c) = E(Y_{11}|A=1,c) = E(Y_{11}|A=0,c) = E(Y_{11}|A=0,M=1,c)$$
$$= E(Y_{11}|A=0,M_0=1,c) = E(Y_{11}|A=0,M_1=M_0=1,c) = E(Y_{11}|M_1=M_0=1,c)$$
$$= E(Y_{1M_1}|M_1=M_0=1,c) = E(Y_1|M_1=M_0=1,c).$$
$$E(Y|A=0,M=1,c) = E(Y_{01}|A=0,M=1,c) = E(Y_{01}|A=0,M_0=1,c) = E(Y_{01}|M_0=1,c)$$
$$= E(Y_{01}|M_1=M_0=1,c) = E(Y_{0M_0}|M_1=M_0=1,c) = E(Y_0|M_1=M_0=1,c).$$

Note that if instead of $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ and $Y_{am} \coprod M|\{A, C\}$ we assume that $\{Y_a, M_{a^*}\} \coprod A|C$ and $Y_a \coprod A|\{M = a, C\}$ then we still have:

$$E(Y|A=1,M=0,c) = E(Y_1|A=1,M=0,c) = E(Y_1|A=1,M_1=0,c) = E(Y_1|M_1=0,c)$$
$$= E(Y_1|M_1=M_0=0,c)$$

where the first and second equalities follow by consistency, the third by the assumption that $\{Y_a, M_{a^*}\} \coprod A|C$ and

the fourth by monotonicity. We also have that

$$E(Y|A=0,M=0,c) = E(Y_0|A=0,M=0,c) = E(Y_0|A=1,M=0,c) = E(Y_0|A=1,M_1=0,c)$$
$$= E(Y_0|M_1=0,c) = E(Y_0|M_1=M_0=0,c)$$

where the first and third equalities follows by consistency, the second by the assumption that $Y_0 \coprod A|\{M=0,C\}$, the fourth by the assumption that $\{Y_a, M_{a^*}\} \coprod A|C$ and the fifth by monotonicity. Similarly,

$$E(Y|A=1,M=1,c) = E(Y_1|A=1,M=1,c) = E(Y_1|A=0,M=1,c) = E(Y_1|A=0,M_0=1,c)$$
$$= E(Y_1|M_0=1,c) = E(Y_1|M_1=M_0=1,c).$$
$$E(Y|A=0,M=1,c) = E(Y_0|A=0,M=1,c) = E(Y_0|A=0,M_0=1,c) = E(Y_0|M_0=1,c)$$
$$= E(Y_0|M_1=M_0=1,c).$$

We have established the identification result under the assumptions $\{Y_{am}, M_a, M_{a^*}\} \coprod A|C$ and $Y_{am} \coprod M|\{A,C\}$ (or alternatively under $\{Y_a, M_{a^*}\} \coprod A|C$ and $Y_a \coprod A|\{M=a,C\}$). Now suppose there is an unmeasured confounding variable such that $\{Y_{am}, M_a, M_{a^*}\} \coprod A|\{C,U\}$ and $Y_{am} \coprod M|\{A,C,U\}$ (or such that $\{Y_a, M_{a^*}\} \coprod A|\{C,U\}$ and $Y_a \coprod A|\{M=a,C,U\}$) then we would have

$$E(Y_1-Y_0|M_1=M_0=m,C=c,U=u) = E(Y|A=1,M=m,C=c,U=u) - E(Y|A=0,M=m,C=c,U=u).$$

By exactly the same derivation in Theorem 3 and Corollary 3 we have for any reference level u' that

$$\begin{aligned} E(Y_1-Y_0|M_1=M_0=m,C=c) \quad = \quad & E(Y|A=1,m,c) - E(Y|A=0,m,c) \\ & - [\textstyle\sum_u \{E[Y|A=1,m,c,u] - E[Y|A=1,m,c,u']\}\{P(u|A=1,m,c) - P(u|c)\} \\ & - \textstyle\sum_u \{E[Y|A=0,m,c,u] - E[Y|A=0,m,c,u']\}\{P(u|A=0,m,c) - P(u|c)\}] \end{aligned}$$

and if U is binary with $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0]$ constant across strata of a,c so that $E[Y|a,m,c,U=1] - E[Y|a,m,c,U=0] = \gamma$ and $P(U=1|A=1,m,c) - P(U=1|A=0,m,c)$ constant across strata of c so that $P(u|A=1,m,c) - P(u|A=0,m,c) = \delta$ then

$$E(Y_1-Y_0|M_1=M_0=m,C=c) = E(Y|A=1,m,c) - E(Y|A=0,m,c) - \delta\gamma.$$

This completes the proof.