

Online supplemental material

Estimating the Causal Effect of an Exposure on Change from Baseline Using Directed Acyclic Graphs and Path Analysis

Benoît Lepage, Sébastien Lamy, Dominique Dedieu, Nicolas Savy, Thierry Lang.

1 Regression to the mean examples

Regression to the mean results from intra-individual variability and measurement error on the baseline outcome value, which creates a negative correlation between the measured baseline value and measured change from baseline. We give two examples of bias that can result from this phenomenon.

1.1 Example 1: The exposure depends on baseline level

Regression to the mean is known to create a spurious correlation between the exposure and change from baseline when the exposure depends on baseline level.

Let us assume a very homogeneous population (with no inter-individual variability) in which systolic blood pressure is measured twice (at time $t_1 = 0$ and $t_2 = 10$). In this population, the measured blood pressure at baseline (t_1) has a mean of 120 mmHg and a standard deviation of 10 mmHg (resulting only from intra-individual variability and measurement error).

Furthermore, let us assume that the exposure of interest ($E = 1$) is only given to individuals with $BP^*(t_1) \geq 120$ mmHg (black circles in Figure S1) and that individuals with $BP^*(t_1) < 120$ mmHg are unexposed ($E = 0$) (white circles in Figure S1).

In the situation of a true null hypothesis (the exposure E has no causal effect), where blood pressure does not change in time except for intra-individual variability and measurement error, the observed blood pressure at time t_2 will also have a mean of 120 mmHg and a standard deviation of 10 mmHg.

Individuals with the highest blood pressure values at time t_1 are more likely to have lower blood pressure values at time t_2 , and individuals with the lowest blood pressure values at time t_1 are more likely to have higher blood pressure values at time t_2 (as shown in Figure S1). In such a situation, we observe a decrease in blood pressure for subjects exposed to $E = 1$ and an increase for unexposed subjects, leading to the erroneous conclusion of a non-null protective effect of the exposure.

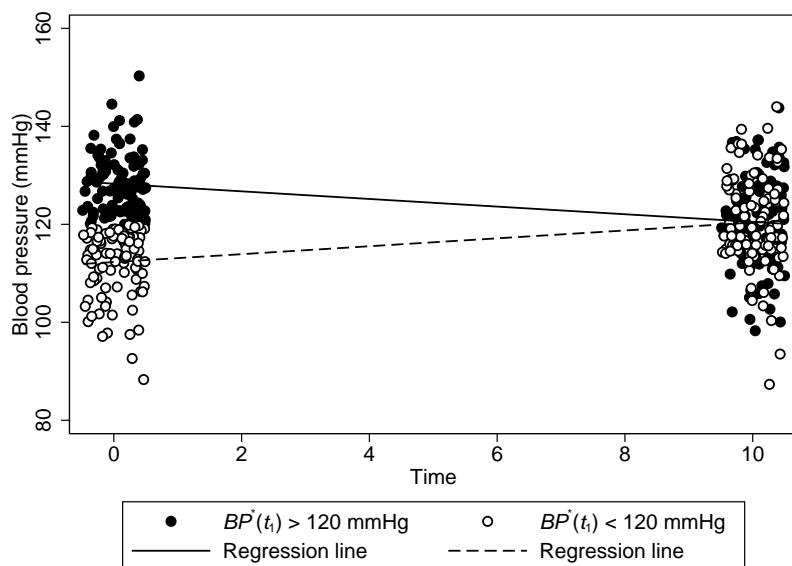


Figure S1: Regression to the mean example, in which the exposure (black circles vs white circles) depends on baseline level

1.2 Example 2: Conditioning on baseline level in pre-existing populations differing by the exposure and baseline level of the outcome

Let us assume two pre-existing homogeneous populations (with no inter-individual variability): the first population is exposed to the exposure of interest ($E = 1$) and have a measured blood pressure at baseline (t_1) of mean 130 mmHg and standard deviation 10 mmHg (resulting only from intra-individual variability and measurement error), the second population is unexposed to E ($E = 0$) and have a measured blood pressure at baseline of mean 110 mmHg and standard deviation 10 mmHg. Two situations can lead to such circumstances:

- studies in which the exposure E starts before the beginning of the study and influences the baseline blood pressure $BP(t_1)$,
- studies with a common causal factor of E and $BP(t_1)$.

In the situation of a true null hypothesis (the exposure E has no causal effect on change), where blood pressure does not change in time except for intra-individual variability and measurement error, the observed blood pressure at time $t_2 = 10$ will also have a mean of 130 mmHg in the first population and a mean of 110 mmHg in the second population.

When conditioning on the value of $BP^*(t_1)$, for example including only subjects with a measured blood pressure at baseline $BP^*(t_1)$ in the interval [115 mmHg; 125 mmHg]:

- the included subjects in the exposed sample are more likely to have higher values at t_2 ,
- the included subjects in the unexposed sample are more likely to have lower values at t_2 ,

leading to the erroneous conclusion of a non-null effect of the exposure on change from baseline (Figure S2).

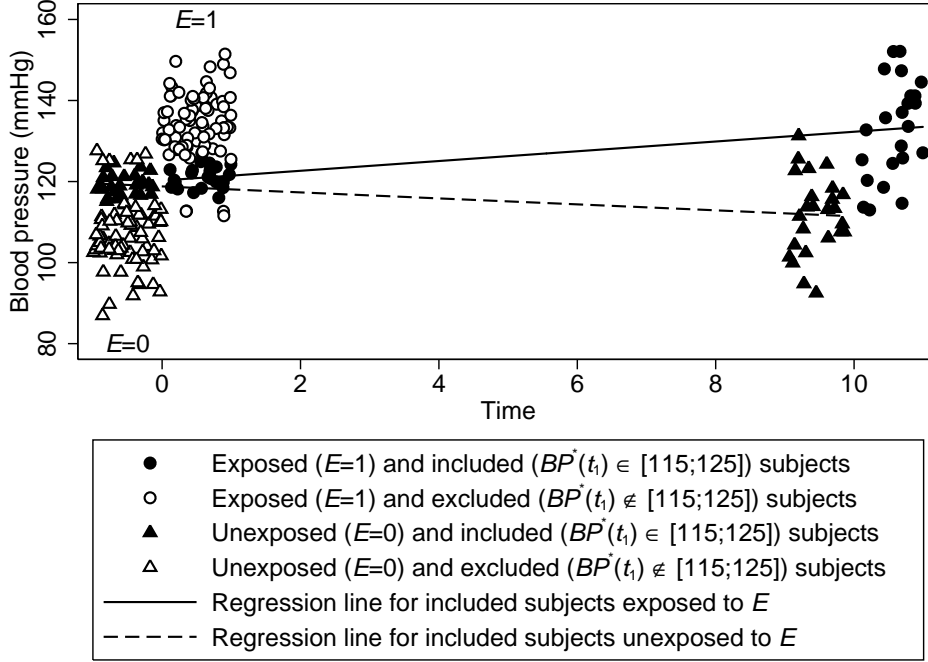


Figure S2: Regression to the mean example, conditioning on baseline level in two pre-existing populations differing by the exposure and baseline level of the outcome

2 Data sets simulation

2.1 General Points

We illustrated the different situations represented in Figure 1 to Figure 4 by simulated data sets compatible with the causal structures of the DAGs. We estimated the effect of the exposure E on blood pressure change ΔBP using linear regression adjusted or unadjusted for the measured baseline level of blood pressure $BP^*(t_1)$. The simulation code for Stata SE 11.2 is provided and can be downloaded online.

In order to simulate a common causal factor of $BP(t_1)$ and ΔBP corresponding to the variable P in Figures 1 to 4, we assumed that blood pressure (BP) varies over time as a polynomial function of age and age^2 , where E is a "correlate of change", so that aging have a varying effect on BP according to the exposure status.[1] With such a polynomial function of age, ΔBP depends on $age(t_1)$ and the effect of E on ΔBP is modified by $age(t_1)$, as shown by Clarke and detailed below.[2] We considered the following parameters to simulate the data sets:

- The length of the study is 10 years. As a consequence, we have $age_i(t_2) = age_i(t_1) + 10$.
- The exposure E is a binary intervention with a mean effect of -10mmHg on blood pressure for exposed versus unexposed individuals 50 years old (except in a few situations indicated below.)
- The age at the beginning of the study, $age_i(t_1)$, was simulated from 40 to 60 years from a uniform distribution (mean age was 50 years). We denote by $c.age(t_1)$ the age centered on the mean age of the sample (at the beginning of the study):

$$c.age_i(t_1) = age_i(t_1) - \overline{age(t_1)} \quad \forall i = 1, \dots, I$$

- To characterize intra-individual variation and measurement error, we simulated the U_{BP1} and U_{BP2} variables independently from a standard Gaussian distribution of variance $\sigma_{U_{BP1}}^2 = \sigma_{U_{BP2}}^2 = \frac{100}{3}$. In Figure 1A, the corresponding intraclass correlation for the measured blood pressure $BP^*(t_1)$ at the beginning of the study is $ICC = \frac{\sigma_{BP^*(t_1)}^2}{\sigma_{BP^*(t_1)}^2 + \sigma_{U_{BP1}}^2} = 0.75$.

We calculated the causal effect of E on ΔBP for individuals aged 50, using the estimated parameters $\hat{\tau}_E^*$ and $\hat{\tau}_E^{**}$ from the following linear regressions (corresponding to models 3 and 4 in the main manuscript). As suggested by Clarke, we included interaction terms $(c.age(t_1) * E)$ and $(c.age(t_1)^2 * E)$ because blood pressure values were simulated according to a quadratic growth curve depending on age and age^2 . [2]

The linear regression adjusted for $BP^*(t_1)$ is (model 3):

$$\mathbb{E}(\Delta BP^* | E, BP^*(t_1), c.age(t_1)) = \mu^* + \tau_{BP1}^* BP^*(t_1) + \tau_E^* E + \tau_{age}^* c.age(t_1) + \tau_{age*E}^* (c.age(t_1) * E) + \tau_{age^2*E}^* (c.age(t_1)^2 * E)$$

The linear regression unadjusted for $BP^*(t_1)$ is (model 4):

$$\mathbb{E}(\Delta BP^* | E, c.age(t_1)) = \mu^{**} + \tau_E^{**} E + \tau_{age}^{**} c.age(t_1) + \tau_{age*E}^{**} (c.age(t_1) * E) + \tau_{age^2*E}^{**} (c.age(t_1)^2 * E)$$

As shown below, the last interaction term $\tau_{age^2*E}^{**} (c.age(t_1)^2 * E)$ is not necessary in the fourth situation where E starts before the beginning of the study and influences both $BP(t_1)$ and $BP(t_2)$ (Figure 4 in the main text).

The directed acyclic graphs in Figures 1 to 4 represent J different scenarios. In each scenario, we simulated $K = 1050$ samples of size $I = 500$. A sample size of 500 was enough to control some instability in the estimations of the effect of E on ΔBP which resulted from applying linear regressions adjusted for $E * age(t_1)$ and $E * [age(t_1)]^2$ interaction terms. Moreover, a sample size of 500 was larger than necessary to detect a 10 mmHg difference of blood pressure between exposed and unexposed subjects in every scenario, with a power greater than 90% and a type I error of 5%. The number of 1050 samples was chosen to be able to explore biases larger than 10% of the standard error of the estimated effect of the exposure E on change ΔBP .

Denoting τ_j^t the “true” causal effect of the exposure E on change ΔBP in the scenario j , and $\hat{\tau}_{jk}$ the effect estimated in the sample k from the linear regression adjusted ($\hat{\tau}_E^*$) or unadjusted for baseline level ($\hat{\tau}_E^{**}$), we calculated: [3]

1. the average bias of the estimated effect of E on ΔBP in the scenario j :

$$bias_j = \bar{\tau}_j - \tau_j^t \quad \text{where} \quad \bar{\tau}_j = \frac{1}{1050} \sum_k \hat{\tau}_{jk}$$

2. the standard error (SE) of the estimated effect of E on ΔBP in the scenario j :

$$SE_j = \sqrt{\frac{1}{1049} \sum_k (\hat{\tau}_{jk} - \bar{\tau}_j)^2}$$

2.2 Randomized Trials

In order to simulate blood pressure values at time t_1 and t_2 for the individual i , we used the following equations:

$$BP_i^{\text{growth}}(\text{age}_i(t_1)) = [\eta_0 + R_{0i}] + [\eta_1 + R_{1i}] (\text{age}_i(t_1)) + [\eta_2 + R_{2i}] (\text{age}_i(t_1))^2 \quad (\text{S1})$$

and

$$\begin{aligned} BP_i^{\text{growth}}(\text{age}_i(t_2), E_i, M_i) = & [\eta_0 + (\theta_0 + \xi_{0i}) E_i + (\delta_0 + \varepsilon_0) M_i + R_{0i}] \\ & + [\eta_1 + (\theta_1 + \xi_{1i}) E_i + R_{1i}] (\text{age}_i(t_1) + 10) \\ & + [\eta_2 + (\theta_2 + \xi_{2i}) E_i + R_{2i}] (\text{age}_i(t_1) + 10)^2 \end{aligned} \quad (\text{S2})$$

where a random half of the population is exposed to E ($E_i = 1$) at the beginning of the study.

Parameters ξ_0, ξ_1, ξ_2 and ε_0 are exogenous and independent random variables (from a Gaussian distribution) used to add inter-individual variability for the effect of the variables E and M .

We denote $t.\text{age}_i(t_1) = \text{age}_i(t_1) - 50$ the age centered on $\mathbb{E}(\text{age}(t_1)) = 50$ in the general population, so that $\mathbb{E}(t.\text{age}(t_1)) = 0$.

According to the equations (S1) and (S2), change from baseline for a given exposure E_i and $\text{age}_i(t_1) = (t.\text{age}_i(t_1) + 50)$ can be written as :

$$\begin{aligned} \Delta BP(\text{age}_i(t_1), E_i, M_i) = & BP_i^{\text{growth}}(\text{age}_i(t_2), E_i, M_i) - BP_i^{\text{growth}}(\text{age}_i(t_1)) \\ \Delta BP(\text{age}_i(t_1), E_i, M_i) = & 10(\eta_1 + R_{1i}) + 1100(\eta_2 + R_{2i}) + (\delta_0 + \varepsilon_{0i}) M_i \\ & + [(\theta_0 + \xi_{0i}) + 60(\theta_1 + \xi_{1i}) + 3600(\theta_2 + \xi_{2i})] E_i \\ & + 20(\eta_2 + R_{2i}) t.\text{age}_i(t_1) \\ & + [(\theta_1 + \xi_{1i}) + 120(\theta_2 + \xi_{2i})] E_i t.\text{age}_i(t_1) \\ & + (\theta_2 + \xi_{2i}) E_i [t.\text{age}_i(t_1)]^2 \end{aligned} \quad (\text{S3})$$

The expected causal effect on change from baseline of $\text{do}(E = 1)$ vs $\text{do}(E = 0)$ for a given $\text{age}(t_1) = (t.\text{age}_i(t_1) + 50)$ is equal to:

$$\begin{aligned} \mathbb{E}(\Delta BP(\text{do}(E = 1), \text{age}(t_1)) - \Delta BP(\text{do}(E = 0), \text{age}(t_1))) = & (\theta_0 + 60\theta_1 + 3600\theta_2) \\ & + (\theta_1 + 120\theta_2) t.\text{age}(t_1) + \theta_2 [t.\text{age}(t_1)]^2 \end{aligned} \quad (\text{S4})$$

We can see in equation (S4) that $\text{age}(t_1)$ is an effect modifier of the causal effect of E on ΔBP (the causal difference is not constant in $\text{age}(t_1)$). [4] In order to test the null hypothesis, one needs to test $\{H_0 : \theta_0 = 0 \text{ and } \theta_1 = 0 \text{ and } \theta_2 = 0\}$, which is equivalent to testing $\{H_0 : \tau_E^* = 0 \text{ and } \tau_{\text{age}^*E}^* = 0 \text{ and } \tau_{\text{age}^2*E}^* = 0\}$ or $\{H_0 : \tau_E^{*'} = 0 \text{ and } \tau_{\text{age}^*E}^{*'} = 0 \text{ and } \tau_{\text{age}^2*E}^{*'} = 0\}$.

A simple way to deal with this effect modification in our examples is to focus on the effect of E on ΔBP for the mean $\text{age}(t_1)$ value, so that interaction terms $(t.\text{age}(t_1) * E)$ and $(t.\text{age}(t_1)^2 * E)$ cancel out, and the null hypothesis $H_0 : \{\theta_0 + 60\theta_1 + 3600\theta_2 = 0\}$ is equivalent to testing $H_0 : \{\tau_E^* = 0\}$ with model (4) or $H_0 : \{\tau_E^{*'} = 0\}$ with model (5).

In Figures 1A and 1B in the main text, we used the following parameters:

- $\eta_0 \approx 121.8$, $\eta_1 \approx 0.0214$ and $\eta_2 \approx 0.00286$ to characterize the general quadratic shape of the BP curve. For individuals unexposed to E and M , these values corresponded to a mean BP of 130 mmHg at 50 years old with a slow quadratic BP increase of +3.4 mmHg between 50 and 60 years old.
- the random variables R_0 , R_1 and R_2 were simulated from Gaussian distributions of respective standard deviation $\sigma_{R_0} = 2$, $\sigma_{R_1} = 0.1$, and $\sigma_{R_2} = 0.001$, to create some inter-individual variability in the general shape of the BP curve
- $\theta_0 = -5$, $\theta_1 = -0.05$ and $\theta_2 = -1/1800$, so that the average effect of the exposure E on ΔBP is equal to -10 mmHg for individuals aged 50
- the random variables ξ_0 , ξ_1 and ξ_2 were simulated from Gaussian distributions of respective standard deviation $\sigma_{\xi_0} = 2$, $\sigma_{\xi_1} = 0.1$ and $\sigma_{\xi_2} = 0.001$, to create inter-individual variability of the exposure effect

In Figure 1B in the main text, we used a binary variable M to simulate some effect of $BP(t_1)$ on $BP(t_2)$:

- for individuals whose $BP(t_1)$ value was higher than 140 mmHg, we simulated the M_i variable according to a random binomial distribution of probability 60%
- for individuals whose $BP(t_1)$ value was lower than 140 mmHg, the variable M was fixed at $M_i = 0$
- $\delta_0 = +5$ mmHg, characterizing the effect of the variable M on $BP(t_2)$
- the random variable ε_0 was simulated from a Gaussian distribution of standard error $\sigma_{\varepsilon_0} = 2$ to add some inter-individual variability for the effect of M on $BP(t_2)$

Relationships between these variables can be represented with additional measurement errors on $BP(t_1)$ and $BP(t_2)$ as in Figure S3.

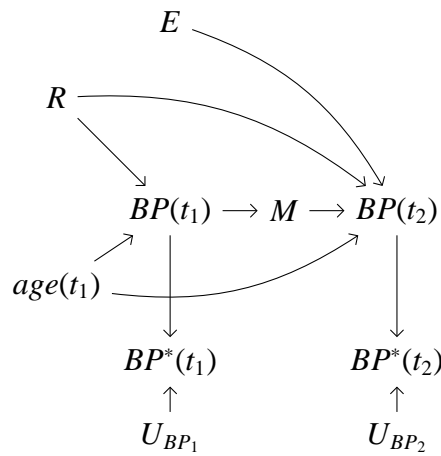


Figure S3: Randomized trials

2.3 Non-randomized Studies with Confounding Factors between the Exposure and the Outcome

To simulate blood pressure values at time t_1 and t_2 for the individual i in the scenarios of Figure 2, we used the following equations:

$$BP_i^{\text{growth}}(age_i(t_1), C_i) = [\eta_0 + (\lambda_0 + \zeta_{0i}) C_i + R_{0i}] + [\eta_1 + (\lambda_1 + \zeta_{1i}) C_i + R_{1i}] (age_i(t_1)) + [\eta_2 + (\lambda_2 + \zeta_{2i}) C_i + R_{2i}] (age_i(t_1))^2 \quad (S5)$$

and

$$BP_i^{\text{growth}}(age_i(t_2), E_i, M_i, C_i) = [\eta_0 + (\theta_0 + \xi_{0i}) E_i + (\delta_0 + \varepsilon_{0i}) M_i + (\lambda_0 + \zeta_{0i}) C_i + R_{0i}] + [\eta_1 + (\theta_1 + \xi_{1i}) E_i + (\lambda_1 + \zeta_{1i}) C_i + R_{1i}] (age_i(t_1) + 10) + [\eta_2 + (\theta_2 + \xi_{2i}) E_i + (\lambda_2 + \zeta_{2i}) C_i + R_{2i}] (age_i(t_1) + 10)^2 \quad (S6)$$

According to equations (S5) and (S6), change from baseline for a given exposure E_i , C_i and $age_i(t_1)$ can be written as :

$$\begin{aligned} \Delta BP(age_i(t_1), E_i, M_i, C_i) &= BP_i^{\text{growth}}(age_i(t_2), E_i, M_i, C_i) - BP_i^{\text{growth}}(age_i(t_1), C_i) \\ \Delta BP(age_i(t_1), E_i, M_i, C_i) &= 10(\eta_1 + R_{1i}) + 100(\eta_2 + R_{2i}) + (\delta_0 + \varepsilon_{0i}) M_i \\ &\quad + [(\theta_0 + \xi_{0i}) + 10(\theta_1 + \xi_{1i}) + 100(\theta_2 + \xi_{2i})] E_i \\ &\quad + [10(\lambda_1 + \zeta_{1i}) + 100(\lambda_2 + \zeta_{2i})] C_i \\ &\quad + 20(\eta_2 + R_{2i}) age_i(t_1) \\ &\quad + [(\theta_1 + \xi_{1i}) + 20(\theta_2 + \xi_{2i})] E_i age_i(t_1) \\ &\quad + 20(\lambda_2 + \zeta_{2i}) C_i age_i(t_1) \\ &\quad + (\theta_2 + \xi_{2i}) E_i [age_i(t_1)]^2 \end{aligned} \quad (S7)$$

The expected causal effect on change from baseline of $do(E = 1)$ vs $do(E = 0)$ for a given $age(t_1)$ is the same as in equation (S4).

We can see in equation (S7) that if $\lambda_1 = \lambda_2 = \zeta_{1i} = \zeta_{2i} = 0$ (*i.e.* when the effect of C on $BP(t_1)$ and $BP(t_2)$ is not modified by $age(t_1)$), then ΔBP does not depend on C so that Figures 2A and 2B can be simplified into Figures 2C and 2D.

We did not simulate data from Figure 2B as it does not provide additional information to the results observed from Figures 2A, 2C and 2D. We defined the following parameters for the simulated data sets:

- C is a binary variable, simulated from a binomial distribution of probability $\mathbb{P}(C = 1) = 0.40$
- C influences the probability of being exposed to E . E was simulated from a conditional binomial distribution of probability

$$\mathbb{P}(E | C) = \text{expit} \left[\ln \left(\frac{0.25}{1 - 0.25} \right) + \ln(3) \times C \right],$$

where $\text{expit}(x) = \exp(x) / [1 + \exp(x)]$, so that $\mathbb{P}(E = 1 | C = 0) = 25\%$ and the effect of C on E is characterized by an odds ratio (OR) of 3.

- C influences blood pressure at the beginning and at the end of the study: individuals exposed to $C = 1$ had a BP $\lambda_0 = 15$ mmHg higher than individuals unexposed to C .
- the random variable ζ_{0i} was simulated from a Gaussian distribution of standard deviation 5 mmHg, adding some inter-individual variability to the effect of C on $BP(t_1)$ and $BP(t_2)$.

In Figures 2A, 2C and 2D in the main text, we used the same parameters as in Figures 1A and 1B regarding: $\eta_0 \approx 121.8$, $\eta_1 \approx 0.0214$ and $\eta_2 \approx 0.00286$, $\theta_0 = -5$, $\theta_1 = -0.05$ and $\theta_2 = -1/1800$. The random variables $R_0, R_1, R_2, \xi_0, \xi_1$ and ξ_2 have been simulated in the same way as in Figures 1A and 1B.

In Figure 2A, the effect of C on $BP(t_1)$ and $BP(t_2)$ is modified by $age(t_1)$, using the following parameters: $\lambda_1 = 0.02$, $\lambda_2 = 0.003$. The random variables ζ_1 and ζ_2 were simulated from Gaussian distributions of respective standard deviation $\sigma_{\zeta_1} = 0.005$, $\sigma_{\zeta_2} = 0.0005$, to create inter-individual variability of the effect modification by $age(t_1)$.

In Figures 2C and 2D, the effect of C on $BP(t_1)$ and $BP(t_2)$ is not modified by $age(t_1)$, so that $\lambda_1 = \lambda_2 = 0$ and the random variables ζ_1 and ζ_2 do not have to be simulated.

In Figure 2D in the main text, the binary variable M and ε_0 have been simulated in the same way as in Figures 1B. The parameter $\delta_0 = +5$ mmHg.

Relationships between these variables can be represented with additional measurement errors on $BP(t_1)$ and $BP(t_2)$ as in Figure S4.

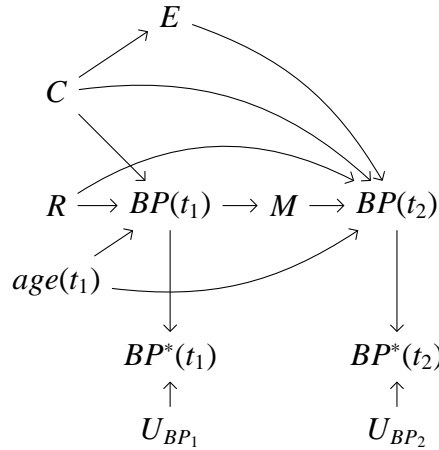


Figure S4: Non-randomized studies with confounders C between the exposure and blood pressure

2.4 Non-randomized Studies where the Observed Baseline Outcome Influences the Exposure

Data set simulation for the DAGs of Figures 3A and 3B

To simulate blood pressure values at time t_1 and t_2 for the individual i in the scenarios of Figure 3, we used the same equations (S1) and (S2) than for the randomized controlled trial data set.

In Figures 3A and 3B in the main text, the observed baseline blood pressure $BP^*(t_1)$ influences the probability of being exposed to E at the beginning of the study. E was simulated from a conditional binomial distribution of probability

$$\mathbb{P}(E \mid BP^*(t_1)) \approx \text{expit} \left[-9.13 + \ln(2^{1/10}) \times BP^*(t_1) \right],$$

so that $\mathbb{P}(E = 1 \mid BP^*(t_1) = 100) = 10\%$ and the odds of $P(E = 1)$ is multiplied by 2 for a 10 mmHg increase of $BP^*(t_1)$.

In Figures 3A and 3B in the main text, we used the same parameters as in Figures 1A and 1B regarding: $\eta_0 \approx 121.8$, $\eta_1 \approx 0.0214$ and $\eta_2 \approx 0.00286$, $\theta_0 = -5$, $\theta_1 = -0.05$ and $\theta_2 = -1/1800$. The random variables R_0 , R_1 , R_2 , ξ_0 , ξ_1 and ξ_2 have been simulated in the same way as in Figures 1A and 1B.

In Figure 3B in the main text, the binary variable M and the variable ε_0 have been simulated in the same way as in Figure 1B. The parameter $\delta_0 = +5$ mmHg.

Relationships between the variables corresponding to Figures 3A and 3B can be represented with additional measurement errors on $BP(t_1)$ and $BP(t_2)$ as in Figure S5.

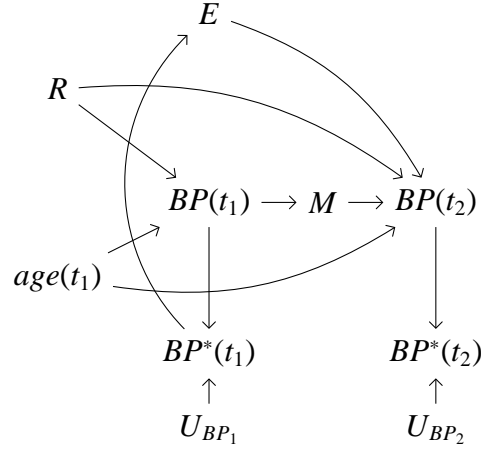


Figure S5: Non-randomized studies where $BP^*(t_1)$ influences the exposure

Data set simulation for the DAG of Figure 3C

In Figure 3C in the main text, the probability of being exposed to E at the beginning of the study is influenced by a pre-existing value of blood pressure $BP^*(t_0)$, which have been measured 10 years earlier ($age(t_0) = age(t_1) - 10$).

In order to simulate blood pressure value at time t_0 for the individual i , we used an equation similar to (S1) with $age(t_0)$ instead of $age(t_1)$:

$$BP_i^{\text{growth}}(age_i(t_0)) = [\eta_0 + R_{0i}] + [\eta_1 + R_{1i}] (age_i(t_0)) + [\eta_2 + R_{2i}] (age_i(t_0))^2 \quad (\text{S8})$$

The exposure E was simulated from a conditional binomial distribution of probability

$$\mathbb{P}(E \mid BP^*(t_0)) \approx \text{expit}[-9.13 + \ln(2^{1/10}) \times BP^*(t_0)],$$

so that $\mathbb{P}(E = 1 \mid BP^*(t_0) = 100) = 10\%$ and the odds of $P(E = 1)$ is multiplied by 2 for a 10 mmHg increase of $BP^*(t_0)$.

Blood pressure values at time t_1 and t_2 for the individual i , have been simulated using the same equations (S1) and (S2) as in Figures 3A and 3B, with the same parameters $\eta_0 \approx 121.8$, $\eta_1 \approx 0.0214$ and $\eta_2 \approx 0.00286$, $\theta_0 = -5$, $\theta_1 = -0.05$ and $\theta_2 = -1/1800$. The random variables R_0 , R_1 , R_2 , ξ_0 , ξ_1 and ξ_2 have also been simulated in the same way as in Figures 3A and 3B.

Relationships between the variables corresponding to Figures 3C can be represented with additional measurement errors on $BP(t_0)$, $BP(t_1)$ and $BP(t_2)$ as in Figure S6.

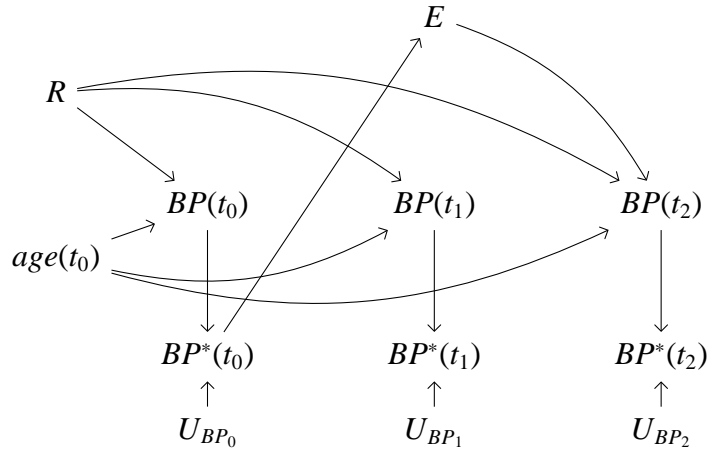


Figure S6: Non-randomized studies where $BP^*(t_0)$ influences the exposure

2.5 Non-randomized Studies where the Exposure Starts before the Beginning of the Study

In DAGs of Figures 4 in the main text, the exposure starts at t_0 (before the beginning of the study). As previously, BP evolves according to a growth curve, so that ΔBP depends on the length of the exposure at time t_1 , $length(t_1) = age(t_1) - age(t_0)$ (i.e. the individual change score depends on both $age(t_0)$ and $age(t_1)$). In order to know the value of $age(t_0)$ for individuals unexposed to $E = 1$, we must assume that the exposure E has to potentially occur at the same $age(t_0)$ for every subject (or at a time origin that can be clearly defined from some common event known for every subject). If we do not know the age at the time of potential exposure for subjects who happened to be unexposed to E , the consistency assumption (the fact that an individual's potential outcome under a hypothetical intervention is the observed outcome if the intervention happened to materialize) does not hold and we cannot estimate the causal effect of E on ΔBP . [5]

To simulate blood pressure values at time t_1 and t_2 for the individual i in the scenarios of Figure 4, we used the following equations:

$$\begin{aligned} BP_i^{\text{growth}}(length_i(t_1), E_i) = & [\eta_0 + (\theta_0 + \xi_{0i}) E_i + R_{0i}] \\ & + [\eta_1 + (\theta_1 + \xi_{1i}) E_i + R_{1i}] (length_i(t_1)) \\ & + [\eta_2 + (\theta_2 + \xi_{2i}) E_i + R_{2i}] (length_i(t_1))^2 \end{aligned} \quad (S9)$$

and

$$\begin{aligned} BP_i^{\text{growth}}(length_i(t_2), E_i, M_i) = & [\eta_0 + (\theta_0 + \xi_{0i}) E_i + (\delta_0 + \varepsilon_0) M_i + R_{0i}] \\ & + [\eta_1 + (\theta_1 + \xi_{1i}) E_i + R_{1i}] (length_i(t_1) + 10) \\ & + [\eta_2 + (\theta_2 + \xi_{2i}) E_i + R_{2i}] (length_i(t_1) + 10)^2 \end{aligned} \quad (S10)$$

where:

- A random half of the population is exposed to E ($E_i = 1$) at the age of 20 years. If, contrary to the fact, unexposed individuals had been exposed, the exposure would have also started at the age of 20 years
- The $length_i(t_1)$ variable is the potential length of the exposure at the beginning (t_1) of the study for the individual i , defined by $length_i(t_1) = age_i(t_1) - 20$
- The potential length of the exposure at the end (t_2) of the study is $length_i(t_2) = length_i(t_1) + 10$

According to equations (S9) and (S10), change from baseline for a given exposure E_i and $length_i(t_1)$ can be written as :

$$\begin{aligned} \Delta BP(length_i(t_1), E_i, M_i) = & BP_i^{\text{growth}}(length_i(t_2), E_i, M_i) - BP_i^{\text{growth}}(length_i(t_1), E_i) \\ \Delta BP(length_i(t_1), E_i, M_i) = & 10(\eta_1 + R_{1i}) + 100(\eta_2 + R_{2i}) + (\delta_0 + \varepsilon_{0i}) M_i \\ & + [10(\theta_1 + \xi_{1i}) + 100(\theta_2 + \xi_{2i})] E_i \\ & + 20(\eta_2 + R_{2i}) length_i(t_1) \\ & + 20(\theta_2 + \xi_{2i}) E_i length_i(t_1) \end{aligned} \quad (S11)$$

It should be noted that in this causal structure, M_i is a function of E_i .

The expected causal effect on change from baseline of $\text{do}(E = 1)$ vs $\text{do}(E = 0)$ for a given $\text{length}(t_1) = (t.\text{age}_i(t_1) + 30)$ is equal to:

$$\begin{aligned} \mathbb{E}(\Delta BP(\text{do}(E = 1), \text{length}(t_1), M_{i,\text{do}(E=1)}) - \Delta BP(\text{do}(E = 0), \text{length}(t_1), M_{i,\text{do}(E=0)})) = \\ 10\theta_1 + 700\theta_2 + 20\theta_2 t.\text{age}(t_1) \\ + \delta_0 [\mathbb{P}(M = 1 | \text{do}(E = 1), \text{length}_i(t_1)) - \mathbb{P}(M = 1 | \text{do}(E = 0), \text{length}_i(t_1))] \quad (\text{S12}) \end{aligned}$$

As previously, $\text{age}(t_1)$ is an effect modifier of the causal risk difference of E on ΔBP (the causal difference is not constant in $\text{age}(t_1)$), and we will focus on the effect of E on ΔBP for the mean $\text{age}(t_1)$ value.

In Figures 4A and 4B, we used the following parameters:

- $\eta_0 = 125$, $\eta_1 = 1/20$ and $\eta_2 = 3/200$ to characterize the general quadratic shape of the BP curve. For individuals unexposed to E and M , these values corresponded to a mean BP of 140 mmHg at age 50 with a slow quadratic BP increase of +10 mmHg between age 50 and 60.
- the random variables R_0 , R_1 and R_2 were simulated from Gaussian distributions of respective standard deviation $\sigma_{R_0} = 2$, $\sigma_{R_1} = 0.1$, and $\sigma_{R_2} = 0.001$, to create some inter-individual variability in the general shape of the BP curve

In Figure 4A:

- $\theta_0 = -5$, $\theta_1 = -3/100$ and $\theta_2 = -97/7000$, so that the average effect of the exposure E on change ΔBP is equal to -10 mmHg for subjects aged 50 at the beginning of the study.
- The random variables ξ_0 , ξ_1 and ξ_2 were simulated from Gaussian distributions of respective standard deviation $\sigma_{\xi_0} = 2$, $\sigma_{\xi_1} = 0.1$ and $\sigma_{\xi_2} = 0.001$, to create inter-individual variability of the exposure effect.

In Figure 4B, we used the same parameters $\theta_0 = -5$, $\theta_1 = -3/100$ and $\theta_2 = -97/7000$ as in Figure 4A and the variables ξ_0 , ξ_1 and ξ_2 have been simulated in the same way.

The binary variable M and the variable ε_0 have been simulated in the same way as in Figure 1B. The effect of M on $BP(t_2)$ was higher than previously with $\delta_0 = +15$ mmHg.

We can calculate from the equation of $BP(t_1)$:

$$\begin{aligned} \mathbb{P}(BP(t_1) > 140 | \text{do}(E = 0), \text{age}(t_1) = 50) &\approx 50\% \\ \text{and} \\ \mathbb{P}(BP(t_1) > 140 | \text{do}(E = 1), \text{age}(t_1) = 50) &\approx 0.024\% \end{aligned}$$

As the binary variable M appears in 60% of the subjects whose $BP(t_1)$ value is higher than 140 mmHg, the “true” effect of E on change ΔBP for individuals aged 50 is:

$$\tau^t = -10 + \delta_0 \times 0.6 \times (0.024 - 0.50) \approx -14.50 \text{ mmHg}$$

Relationships between these variables can be represented with additional measurement errors on $BP(t_1)$ and $BP(t_2)$ as in Figure S7.

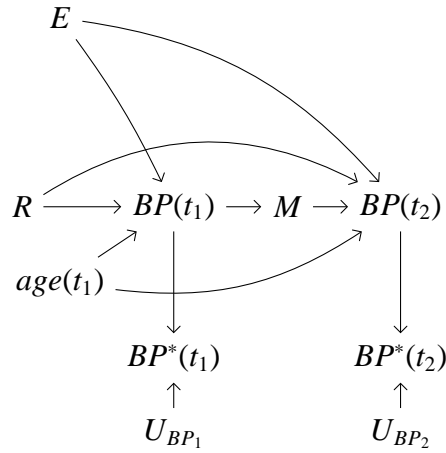


Figure S7: Non-randomized studies where E starts before the beginning of the study

References

- [1] Rogosa DR, Willett JB. Understanding correlates of change by modeling individual difference in growth. *Psychometrika*. 1985;50(2):203–228.
- [2] Clarke PS. Causal analysis of individual change using the difference score. *Epidemiology*. 2004;15(4):414–421.
- [3] Burton A, Altman DG, Royston P, Holder RL. The design of simulation studies in medical statistics. *Statistics in Medicine*. 2006;25(24):4279–92.
- [4] VanderWeele TJ, Robins JM. Four types of effect modification: a classification based on directed acyclic graphs. *Epidemiology*. 2007;18(5):561–568.
- [5] Pearl J. On the consistency rule in causal inference. Axiom, definition, assumption, or theorem? *Epidemiology*. 2010;21(6):872–875.

```

***** Simulation code for Stata/SE 11.2
***** Estimating the causal effect of an exposure on change from baseline using DAGs and path analysis
***** 2014 june

set mem lg

program linear_regressions_of_change
    version ll
    args BP1_mes BP2_mes E age1
    sum `age1'
    scalar mean_age = r(mean)
    gen age_c = `age1' - mean_age
    gen age_c2 = age_c^2

    gen dif = `BP2_mes' - `BP1_mes'

    * model 4) linear regression adjusted on BP1 +/- E*age and E*age^2 interaction terms
    xi: regress dif `E' `BP1_mes' age_c i.`E'*age_c i.`E'*age_c2
    scalar adj_lr_age2 = _b[`E']

    xi: regress dif `E' `BP1_mes' age_c i.`E'*age_c
    scalar adj_lr_age = _b[`E']

    xi: regress dif `E' `BP1_mes'
    scalar adj_lr = _b[`E']

    * model 5) linear regression unadjusted on BP1 +/- E*age and E*age^2 interaction terms
    xi: regress dif `E' age_c i.`E'*age_c i.`E'*age_c2
    scalar unadj_lr_age2 = _b[`E']

    xi: regress dif `E' age_c i.`E'*age_c
    scalar unadj_lr_age = _b[`E']

    xi: regress dif `E'
    scalar unadj_lr = _b[`E']

    drop age_c age_c2 dif
end

*****
*** I) Randomized control trial datasets simulation
*****

*****
*** 1A) Figure 1A
clear
set seed X64ad479b45c155c3ddda364fd8449359000243b6
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

```

```

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen BP1 = [eta0 + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen E = rbinomial(1,0.5)

gen BP2 = [eta0 + (theta0 + xi0)*E + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_1a.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_1a.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_1a_`i'.dta", replace

```



```

}

use "D:\...simulation_file_path...\results_sim_1a_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_1a_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_1a.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_1a_`i'.dta"
}

*** Results simulation Figure 1A
use "D:\...simulation_file_path...\results_sim_1a.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

*****
*** 1B) Figure 1B
clear
set seed Xf73e2f9f2b0a423625c79d6ba81d8ec800024cbe
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen BP1 = [eta0 + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

gen M = rbinomial(1,0.6) if BP1>=140
replace M = 0 if BP1<140
gen delta0 = 5
gen eps0=rnormal(0,2)

gen E = rbinomial(1,0.5)

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)

```

```

gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen BP2 = [eta0 + (theta0 + xi0)*E + (delta0 + eps0)*M + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_1b.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_1b.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_1b_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_1b_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_1b_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_1b.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_1b_`i'.dta"
}

*** Results simulation Figure 1B
use "D:\...simulation_file_path...\results_sim_1b.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

```

```

*****
*** II) Non-randomized Studies with Confounding Factors between the Exposure and the Outcome
*****

*****
*** 2A) Figure 2A
clear
set seed X8a8d036b45674c0eec8a50a508fed9f500020f2a
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

gen C = rbinomial(1,0.40)
gen E = rbinomial(1, invlogit( ln(0.25/(1-0.25)) + C*ln(3)) )

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen lambda0 = 15
gen zeta0 = rnormal(0,5)

*gen lambda1 = 0.01
gen lambda1 = 0.02
gen zeta1 = rnormal(0,0.005)
*gen lambda2 = 0.001
gen lambda2 = 0.003
gen zeta2 = rnormal(0,0.0005)

gen BP1 = [eta0 + (lambda0 + zeta0)*C + R0] + [eta1 + (lambda1 + zeta1)*C + R1]*age1 + [eta2 + (lambda2 + zeta2)*C + R2]*(age1^2)

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen BP2 = [eta0 + (theta0 + xi0)*E + (lambda0 + zeta0)*C + R0] + [eta1 + (theta1 + xi1)*E + (lambda1 + zeta1)*C + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + (lambda2 + zeta2)*C + R2]*((age1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

```

```

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_2a.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_2a.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen effect_adj=scalar(adj_lr)
gen effect_unadj=scalar(unadj_lr)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen bias_adj = effect_adj + 10
gen bias_unadj = effect_unadj + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_2a_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_2a_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_2a_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_2a.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_2a_`i'.dta"
}

*** Results simulation Figure 2A
use "D:\...simulation_file_path...\results_sim_2a.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

```

```

*****
*** 2C) Figure 2C
clear
set seed Xbac0c56195c254508582d28d7a2e469100022645
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

gen C = rbinomial(1,0.40)
gen E = rbinomial(1, invlogit( ln(0.25/(1-0.25)) + C*ln(3)) )

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen lambda0 = 15
gen zeta0 = rnormal(0,5)

gen BP1 = [eta0 + (lambda0 + zeta0)*C + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen BP2 = [eta0 + (theta0 + xi0)*E + (lambda0 + zeta0)*C + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_2c.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_2c.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

```

```

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen effect_adj=scalar(adj_lr)
gen effect_unadj=scalar(unadj_lr)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen bias_adj = effect_adj + 10
gen bias_unadj = effect_unadj + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_2c_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_2c_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_2c_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_2c.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_2c_`i'.dta"
}

*** Results simulation Figure 2C
use "D:\...simulation_file_path...\results_sim_2c.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

*****
*** 2D) Figure 2D
clear
set seed X505828a031e9171c1b7dcd60f31936c300022033
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

gen C = rbinomial(1,0.40)
gen E = rbinomial(1, invlogit( ln(0.25/(1-0.25)) + C*ln(3)) )

```

```

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen lambda0 = 15
gen zeta0 = rnormal(0,5)

gen BP1 = [eta0 + (lambda0 + zeta0)*C + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen M = rbinomial(1,0.6) if BP1>=140
replace M = 0 if BP1<140
gen delta0 = 5
gen eps0=rnormal(0,2)

gen BP2 = [eta0 + (theta0 + xi0)*E + (delta0 + eps0)*M + (lambda0 + zeta0)*C + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

save "D:\...simulation_file_path...\sim_2d.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_2d.dta", clear
drop if n_simu!=`i'
keep n_simu id BP1_mes BP2_mes E age1

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

```

```

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...\simulation_file_path...\results_sim_2d_`i'.dta", replace
}

use "D:\...\simulation_file_path...\results_sim_2d_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...\simulation_file_path...\results_sim_2d_`i'.dta"
}
save "D:\...\simulation_file_path...\results_sim_2d.dta", replace

forvalues i = 1/1050 {
erase "D:\...\simulation_file_path...\results_sim_2d_`i'.dta"
}

*** Results simulation Figure 2D
use "D:\...\simulation_file_path...\results_sim_2d.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

```

```

*****
*** III) Non-randomized Studies where the Observed Baseline Outcome Influences the Exposure
*****

```

```

*****
*** 3A) Figure 3A

```

```

clear
set seed Xa8e910a73f9ebf951ff8adeee201b13800022ee1
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen BP1 = [eta0 + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

```



```

gen Ubp1 =rnormal(0,5.7735027)
gen BP1_mes = BP1 + Ubp1

gen E = rbinomial(1,invlogit((logit(0.10) - (ln(2^0.1)*100)) + BP1_mes*ln(2^0.1) ) )

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen BP2 = [eta0 + (theta0 + xi0)*E + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp2 =rnormal(0,5.7735027)
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_3a.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_3a.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_3a_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_3a_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_3a_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_3a.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_3a_`i'.dta"
}

```

```
}
```

```
*** Results simulation Figure 3A
```

```
use "D:\...\simulation_file_path...\results_sim_3a.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)
```

```
*****
```

```
*** 3B) Figure 3B
```

```
clear
```

```
set seed X1546cb3eae93f0add3fc4d9a993e791900021ad1
```

```
set obs 525000
```

```
gen id = _n
```

```
gen n_simu=int( ( (id-1)/500) + 1 )
```

```
gen age1 = 40 + (20*runiform())
```

```
gen eta0 = 121.78571
```

```
gen eta1 = 0.0214288
```

```
gen eta2 = 0.00285714
```

```
gen R0 = rnormal(0,2)
```

```
gen R1 = rnormal(0,0.1)
```

```
gen R2 = rnormal(0,0.001)
```

```
gen BP1 = [eta0 + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)
```

```
gen Ubp1 =rnormal(0,5.7735027)
```

```
gen BP1_mes = BP1 + Ubp1
```

```
gen M = rbinomial(1,0.6) if BP1>=140
```

```
replace M = 0 if BP1<140
```

```
gen delta0 = 5
```

```
gen eps0=rnormal(0,2)
```

```
gen E = rbinomial(1,invlogit((logit(0.10) - (ln(2^0.1)*100)) + BP1_mes*ln(2^0.1) ) )
```

```
gen theta0 = -5
```

```
gen theta1 = -0.05
```

```
gen theta2 = -1/1800
```

```
gen xi0 = rnormal(0,2)
```

```
gen xi1 = rnormal(0,0.1)
```

```
gen xi2 = rnormal(0,0.001)
```

```
gen BP2 = [eta0 + (theta0 + xi0)*E + (delta0 + eps0)*M + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)
```

```
gen Ubp2 =rnormal(0,5.7735027)
```

```

gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_3b.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_3b.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...simulation_file_path...\results_sim_3b_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_3b_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_3b_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_3b.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_3b_`i'.dta"
}

*** Results simulation Figure 3B
use "D:\...simulation_file_path...\results_sim_3b.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

*****
*** 3C) Figure 3C
clear
set seed X18f05abd3ce62fe6ae24356b77ee2b6300024d8a
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

```

```

gen age1 = 40 + (20*runiform())
gen age0 = age1 - 10

gen eta0 = 121.78571
gen eta1 = 0.0214288
gen eta2 = 0.00285714

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen BP0 = [eta0 + R0] + [eta1 + R1]*age0 + [eta2 + R2]*(age0^2)
gen BP1 = [eta0 + R0] + [eta1 + R1]*age1 + [eta2 + R2]*(age1^2)

gen Ubp0 =rnormal(0,5.7735027)
gen BP0_mes = BP0 + Ubp0

gen Ubp1 =rnormal(0,5.7735027)
gen BP1_mes = BP1 + Ubp1

gen E = rbinomial(1,invlogit((logit(0.10) - (ln(2^0.1)*100)) + BP0_mes*ln(2^0.1) ) )

gen theta0 = -5
gen theta1 = -0.05
gen theta2 = -1/1800

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen BP2 = [eta0 + (theta0 + xi0)*E + R0] + [eta1 + (theta1 + xi1)*E + R1]*(age1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((age1 + 10)^2)

gen Ubp2 =rnormal(0,5.7735027)
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP0_mes BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_3c.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_3c.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age2=scalar(adj_lr_age2)
gen effect_unadj_age2=scalar(unadj_lr_age2)

gen bias_adj_age2 = effect_adj_age2 + 10
gen bias_unadj_age2 = effect_unadj_age2 + 10

```

```

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age2 effect_unadj_age2 bias_adj_age2 bias_unadj_age2

save "D:\...\simulation_file_path...\results_sim_3c_`i'.dta", replace
}

use "D:\...\simulation_file_path...\results_sim_3c_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...\simulation_file_path...\results_sim_3c_`i'.dta"
}
save "D:\...\simulation_file_path...\results_sim_3c.dta", replace

forvalues i = 1/1050 {
erase "D:\...\simulation_file_path...\results_sim_3c_`i'.dta"
}

*** Results simulation Figure 3C
use "D:\...\simulation_file_path...\results_sim_3c.dta", clear
tabstat bias_adj_age2 bias_unadj_age2, statistics(mean sd) columns(statistics)
tabstat effect_adj_age2 effect_unadj_age2, statistics(mean sd) columns(statistics)

```

```

*****
*** IV) Non-randomized Studies where the Exposure Starts Before the Beginning of the Study
*****

```

```

*****
*** 4A) Figure 4A
clear
set seed X0346fb596e19cde3d79bd25d550949ee00024b20
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())
gen length1 = age1 - 20

gen eta0 = 125
gen eta1 = 1/20
gen eta2 = 3/200

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

```

```

gen theta0 = -5
gen theta1 = -3/100
gen theta2 = -97/7000

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen E = rbinomial(1,0.5)

gen BP1 = [eta0 + (theta0 + xi0)*E + R0] + [eta1 + (theta1 + xi1)*E + R1]*length1 + [eta2 + (theta2 + xi2)*E + R2]*(length1^2)

gen BP2 = [eta0 + (theta0 + xi0)*E + R0] + [eta1 + (theta1 + xi1)*E + R1]*(length1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((length1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)
gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_4a.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_4a.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age=scalar(adj_lr_age)
gen effect_unadj_age=scalar(unadj_lr_age)

gen bias_adj_age = effect_adj_age - (-10)
gen bias_unadj_age = effect_unadj_age - (-10)

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age effect_unadj_age bias_adj_age bias_unadj_age

save "D:\...simulation_file_path...\results_sim_4a_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_4a_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_4a_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_4a.dta", replace

```

```

forvalues i = 1/1050 {
erase "D:\...\simulation_file_path...\results_sim_4a_`i'.dta"
}

*** Results simulation Figure 4A
use "D:\...\simulation_file_path...\results_sim_4a.dta", clear
tabstat bias_adj_age bias_unadj_age, statistics(mean sd) columns(statistics)
tabstat effect_adj_age effect_unadj_age, statistics(mean sd) columns(statistics)

*****
*** 4B) Figure 4B
clear
set seed Xabdf30f08f5f321c6c1ce432bbceeebc00024cc9
set obs 525000
gen id = _n
gen n_simu=int( ( (id-1)/500) + 1 )

gen age1 = 40 + (20*runiform())
gen length1 = age1 - 20

gen eta0 = 125
gen eta1 = 1/20
gen eta2 = 3/200

gen R0 = rnormal(0,2)
gen R1 = rnormal(0,0.1)
gen R2 = rnormal(0,0.001)

gen theta0 = -5
gen theta1 = -3/100
gen theta2 = -97/7000

gen xi0 = rnormal(0,2)
gen xi1 = rnormal(0,0.1)
gen xi2 = rnormal(0,0.001)

gen E = rbinomial(1,0.5)

gen BP1 = (eta0 + (theta0 + xi0)*E + R0) + (eta1 + (theta1 + xi1)*E + R1)*length1 + (eta2 + (theta2 + xi2)*E + R2)*(length1^2)

gen M = rbinomial(1,0.6) if BP1>=140
replace M = 0 if BP1<140
gen delta0 = 15
gen eps0=rnormal(0,2)

gen BP2 = [eta0 + (theta0 + xi0)*E +(delta0 + eps0)*M + R0] + [eta1 + (theta1 + xi1)*E + R1]*(length1 + 10) ///
          + [eta2 + (theta2 + xi2)*E + R2]*((length1 + 10)^2)

gen Ubp1 =rnormal(0,5.7735027)

```

```

gen Ubp2 =rnormal(0,5.7735027)

gen BP1_mes = BP1 + Ubp1
gen BP2_mes = BP2 + Ubp2

keep n_simu id BP1_mes BP2_mes E age1
save "D:\...simulation_file_path...\sim_4b.dta", replace

*** Analyses
forvalues i = 1/1050 {
use "D:\...simulation_file_path...\sim_4b.dta", clear
drop if n_simu!=`i'

linear_regressions_of_change BP1_mes BP2_mes E age1

gen effect_adj_age=scalar(adj_lr_age)
gen effect_unadj_age=scalar(unadj_lr_age)

gen true_effect = -10 + 15*0.6*[(1 - normal( (140-(125-5+(0.05-3/100)*30+(0.015-97/7000)*(30^2)) )/ sqrt((2^2+2^2) +
(30^2)*(0.1^2+0.1^2)+(30^4)*(0.001^2+0.001^2) ))) //
- (1 - normal( (140-(125+0.05*30+0.015*(30^2)))/sqrt((2^2) + (30^2)*(0.1^2)+(30^4)*(0.001^2) ) ) ) ]

gen bias_adj_age = effect_adj_age - true_effect
gen bias_unadj_age = effect_unadj_age - true_effect

gen id_bis=id - ((`i'-1)*500)

keep if id_bis==1
keep n_simu effect_adj_age effect_unadj_age bias_adj_age bias_unadj_age

save "D:\...simulation_file_path...\results_sim_4b_`i'.dta", replace
}

use "D:\...simulation_file_path...\results_sim_4b_1.dta", clear

forvalues i = 2/1050 {
append using "D:\...simulation_file_path...\results_sim_4b_`i'.dta"
}
save "D:\...simulation_file_path...\results_sim_4b.dta", replace

forvalues i = 1/1050 {
erase "D:\...simulation_file_path...\results_sim_4b_`i'.dta"
}

*** Results simulation Figure 4B
use "D:\...simulation_file_path...\results_sim_4b.dta", clear
tabstat bias_adj_age bias_unadj_age, statistics(mean sd) columns(statistics)
tabstat effect_adj_age effect_unadj_age, statistics(mean sd) columns(statistics)

```