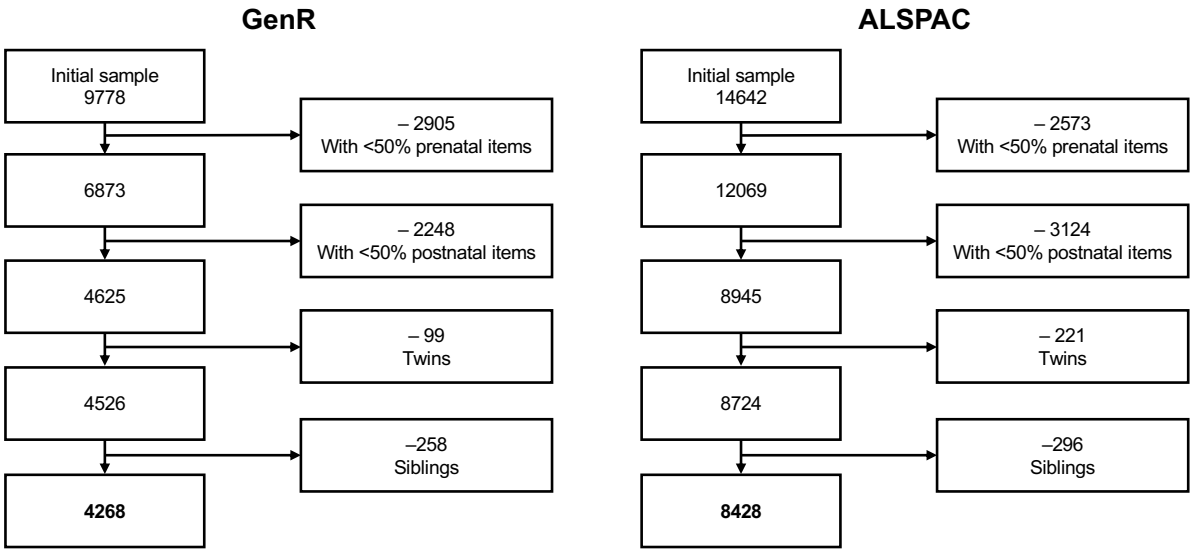
Supplementary Materials

"Effects of pre- and postnatal early-life stress on internalizing, adiposity and their comorbidity"

Table of Contents

Figure S1 | Flowcharts



Flowchart detailing the exclusion steps performed in Generation R (GenR) and ALSPAC.

Supplement 1 | ELS score report

This report describes the development and characteristics of the prenatal and postnatal early-life stress (ELS) scores. These are global ELS measures, intended to capture exposure to cumulative psycho-social risk during fetal life and across childhood (i.e., until the age of 10 years) respectively. The scores were constructed based on data from Generation R (https://generationr.nl/) and ALSPAC (http://www.bristol.ac.uk/alspac/) cohorts. The contents of the following report are publicly available, together with the R scripts used to generate the scores (*[GitHub-link]*).

Score structure

Prenatal and postnatal stressors included in the scores were categorized into five stress *domains,* based on previous literature[1,2]:

- *Life events* (e.g., death of a parent or pregnancy complications),
- *Contextual risk* (e.g., financial difficulties or neighbourhood problems),
- *Parental risk* (e.g., parental criminal record or parental depression),
- *Interpersonal risk* (e.g., family conflicts or loss of a friend),
- *Direct victimization* (only available postnatally, e.g., bullying or maltreatment).

Please see Figure1.png for an overview of all the items included in each stress domain. *Figure 1* (in manuscript) also provides an illustration of the temporal structure of the score, i.e., when were the items measured and what is the period of time they refer to.

Score calculation

In each cohort (i.e., Generation R and ALSPAC), ~100 stress-related items were selected and dichotomised into *no risk* (=0) or *risk* (=1). The dichotomized items were assigned one of 9 stress domains (4 prenatal and 5 postnatal domains, as depicted in *Figure 1*) and a *domain score* was computed as the unweighted average of the items belonging to each domain:

$$\text{domain score} = \text{risk}_1 + \text{risk}_2 + \ldots + \text{risk}_N \,/\, N \text{ (where N is the number of items in the score)}$$

Finally, domain scores were summed within periods to obtain the prenatal and postnatal ELS scores used in the main analyses:

- *Prenatal stress* = prenatal life events + prenatal contextual risk + prenatal parental risk + prenatal interpersonal risk.
- *Postnatal stress* = postnatal life events + postnatal contextual risk + postnatal parental risk + postnatal interpersonal risk + postnatal direct victimization.

Item selection

The score is designed to be a comprehensive measure of exposure to psycho-social stressors. We adopted a broad definition of ELS, including any event or situation with characteristics that would be considered stressful during early life [3]. However, early stress factors that are more biological in nature (e.g. maternal smoking or alcohol consumption, pollution) were left out of the score. We preferred including these as covariates in our models, to explicitly account for their contribution.

Generally, items measured in the prenatal period were referred to the mother, i.e., they were selected to reflect any source of psychosocial stress for the pregnant mother, whereas postnatal items were meant to capture psychosocial stress in the child more directly. However, for a few postnatal ALSPAC items for which more "direct" assessment of child stressors was not available, some proximal items were used. These were referred to the mother, but presumably reflect the child experience to some extent, for example, whether the child's parent died was not reported in ALSPAC, so the item: "the mother's partner died" was used instead. A summary of the overlap between prenatal and postnatal items in each domain and for the two samples is provided here: Table1.png.

When multiformat reports were available (i.e., reports from both parents, or mother and teacher) both sources of information were included, with the aim of reducing the impact of reporter bias as much as possible.

After a first screening of all stress-related questions available in each cohort, we collapsed highly similar/correlated questions to minimize multicollinearity and, at the same time, maximize comparability across cohorts. For example, in the ALSPAC score, the stressor "violence in family" was considered present if the answer was yes to either of these three questions: "Partner was physically cruel to Mum", "Mum/Partner hit or slapped one another" and "Mum/Partner threw something in anger". A complete report of the original questions and their combination into stress items can be downloaded from Table3.xlsx (prenatal items) and Table4.xlsx (postnatal items).

This process resulted in a total of 52 prenatal and 51 postnatal items in Generation R, and 45 prenatal and 50 postnatal items included in the ALSPAC score (see *Figure 1*).

Dichotomization stratergies

To ensure that all different risk items would contribute equally to the score, all items were recoded into *no risk* (=0) or *risk* (=1). About half of the Generation R items and 10% of ALSPAC items were already encoded as binary (i.e., yes or no), but for the remainder of the items decisions about what constitutes a risk (=1) or not (=0) had to be involved. The dichotomization strategy adopted for each item is reported in Table 1.3 and 1.4. In most cases, decisions were guided by questionnaire manuals, European or national guidelines or available literature. Some of the main criteria are outlined here:

- About 65% of all ALSPAC items had a similar 5-level structure: "did not happen" / "no effect at all" / "mildly affected" / "fairly affected" / "affected a lot". These were dichotomized into no risk (=0) if the answer was "did not happen" and risk (=1) if any other answer was provided. This approach would ensure that the subjective evaluation of the impact of a stressor on the child (reported in most cases by mothers) would not factor into the score. The same principle was applied in Generation R, for instance, to questions about financial, housing or interpersonal problems (no risk = "no problems", risk = "slight", "moderate" or "serious" problems), or maternal criminal offence (no risk = "never", risk = "once" / "2-3 times" / "4-5 times" / "> 6 times").
- In the Dutch cohort (Generation R), low education and low income were defined according to the Centraal Bureau voor de Statistiek (2016). Paternal and maternal education was marked as a risk when the highest educational attainment was below "higher education – phase 1" (i.e., risk = No education / Primary / Secondary-phase 1 / Secondary-phase 2). Low income was marked as a risk the household income was below the "basic needs" level (i.e., 1600 €/month).
- The material deprivation criterium was based on a possession rate < 75%, according to European statistics on income and living conditions (EU-SILC, 2007).
- In Generation R, maternal and paternal psychopathology (i.e., depression, anxiety and interpersonal sensitivity) were measured using the Brief Symptom Inventory (BSI). Hence the sex and subscale specific cut-offs were indicated in the Dutch BSI manual[4]. Similarly, in ALSPAC, the manuals of the Edinburgh Postnatal Depression Scale (EPDS)[5] and the Crown-Crisp Experiential Index (CCEI)[6] were consulted to determine clinical cut-offs for depression and anxiety respectively.
- The cut-off for unhealthy family functioning, measured postnatally in Generation R using the Family Assessment Device (FAD) was based on[7].
- Early parenting was defined as a risk when the mother or father was younger than 19 years old at intake (Generation R) or delivery (ALSPAC), based on[1,2].
- Overcrowding (i.e., family size) was included as a risk if more than four people were living in the same house, based on[1,2,8].

- Bullying was considered a risk when any type of bullying (i.e., physical, verbal or relational) was reported by any informant (i.e., mother or teacher) at least once a week, based on[9].
- When no other cut-off was available, an 80th percentile cut-off was used in the dichotomization of maternal and paternal interpersonal sensitivity, and neighborhood problems in ALSPAC and that of maternal and paternal harsh parenting in Generation R, as suggested by[10].

Repeated measures

As evident from the right pane in *Figure 1* (listing postnatal exposures), the two cohorts show different profiles in terms of availability of repeated measurements. In Generation R, none of the prenatal and 16 of the 51 postnatal items were measured more than once:

- Education of the mother and of partner were both measured twice (at ~3 and ~6 years);
- Financial difficulties, trouble paying bills and psychopathology of the mother and partner (i.e., depression, anxiety and interpersonal sensitivity) were measured twice (at ~3 and ~9½ years);
- Bullying was measured when the child was ~6 and ~8 years old;
- Mothers reported on general family functioning when the child was ~6 and ~9½ years old;
- Income, employment status, marital status of the mother and family size were each measured 3 times (at ~3, ~6 and ~9½ years);
- Repeating a grade was measured 3 times (at ~8 years and twice at 9½ – 10 years).

In ALSPAC, while only two prenatal items had repeated measurements (i.e., maternal depression and anxiety were measured in the second and third trimester), the majority of the postnatal stressors had 6 or 7 repeated assessments. The only exceptions being:

- Maternal anxiety and child starting nursery (for which only 5 measurements were available);
- Housing conditions and neighborhood problems, measured twice around ~2 and ~3/4 years;
- Single measurements for: losing a best friend (~8½ years), mother and partner education (~5 years), mother and partner "early parenthood" (at delivery) and bullying (~8 years).

To best leverage data characteristics of each cohort, we therefore decided to adopt a different strategy with respect to handling repeated measurements of postnatal stressors, in Generation R and in ALSPAC.

In Generation R postnatal repeated measures were combined using a '*once a risk, always a risk*' strategy. For example, repeated a grade was encoded as 1 (=risk), whenever a stressor was present at *any* of the available timepoint, and it was encoded as 0 (=no risk) if the stressor was absent at *all* timepoints. There were two exceptions to this rule:

(i) low income and unemployment we additionally combined into "chronic" low income and unemployment items. These were coded as 1 when the stressor was present on *all* timepoints, and as 0 if the stressor was absent at *any* of the three timepoints. This was done because we believe the exposure to an unstable financial situation that is prolonged in time may be a different type of stress exposure. This also implies that low income and unemployment, when experienced chronically, had a bigger impact on the total contextual risk score.

(ii) the two assessments of family functioning according to mothers (at ~6 and ~9½ years) were included in the score as separate stress indicators. Following as similar rationale as for the income and employment variables, we decided not to combine the two measurements, to capture the greater impact that a prolonged exposure to unhealthy family dynamics might exert on interpersonal stress.

In ALSPAC, we believe that adopting the same approach would have been wasteful, considered the richness of postnatal repeated assessments that the cohort provides (i.e., a total of 288 measures across the 50 stressors). Therefore, we did not collapse postnatal repeated assessments using a '*once a risk, always a risk*' strategy but rather we summed all available timepoints as individual items in the domain score, and then divided this sum by the number of stressors in the score:

$$\text{i.e., domain score} = risk_{11} + risk_{12} + \ldots + risk_{1t1} + risk_{21} + \ldots + risk_{2t2} + \ldots + risk_{NtN} / N$$

where N is the number of stressors in the score and t_n is the total number of timepoints available for each stressor. As a result, while in Generation R higher postnatal stress scores reflect mainly the co-occurrence of multiple types of stressors, ALSPAC postnatal scores also reflect the chronicity of stress exposure.

Domain scores distribution and correlational structure

The univariate distributions of each of the nine domains are illustrated here Figure3.png. From the graphs one can appreciate a positive skewness across nearly all domains, that is expected given the population-based samples these distributions are constructed on, i.e., a large proportion of the sample will have low cumulative stress scores. This skewness seems less pronounced in the ALSPAC sample, which could indicate higher rates of cumulative stress. For the prevalence of each individual stressor in the score, see Table3.xlsx (prenatal items) and Table4.xlsx (postnatal items).

The Pearson correlation matrix relating all domain scores as well as prenatal and postnatal total stress scores is shown here: Figure4.jpg. Of note is the high correlation between pre- and postnatal total scores (r = 0.56 in Generation R and 0.48 in ALSPAC), and the Generation R-specific continuity that emerges between prenatal and postnatal contextual risk (r = 0.65). In ALSPAC contextual risk and parental risk also show some continuity across prenatal and postnatal periods (r = 0.40 and 0.43 respectively).

The partitions of individual stressors into the proposed domain structure received some support from the Confirmatory Factor Analyses implemented in the Generation R sample:

Prenatal domains: Life events (RMSEA = 0.03; SRMR = 0.04; CFI = 0.71); Contextual risk (RMSEA = 0.09; SRMR = 0.07; CFI = 0.87); Parental risk (RMSEA = 0.04; SRMR = 0.09; CFI = 0.80); Interpersonal risk (RMSEA = 0.03; SRMR = 0.06; CFI = 0.93).

Postnatal domains: Life events (RMSEA = 0.04; SRMR = 0.04; CFI = 0.49); Contextual risk (RMSEA = 0.07; SRMR = 0.07; CFI = 0.92); Parental risk (RMSEA = 0.05; SRMR = 0.11; CFI = 0.81); Interpersonal risk (RMSEA = 0.05; SRMR = 0.07; CFI = 0.90); Direct victimization (RMSEA = 0.06; SRMR = 0.08; CFI = 0.46).

Why a cumulative ELS score?
Cumulative scores, similar to this one, are widely used in developmental psychology and medicine because they proved to be a parsimonious and statistically sensitive metric. They make no assumptions about the relative strengths of multiple risk factors or their collinearity, and they tend to fit well with underlying theoretical models[11].

Limitations
Additive indexes such as this one also come with a number of shortcomings. Importantly, risk was inevitably designated with some degree of arbitrariness, and dichotomization into risk or no risk was not always a straightforward decision. Because we are aware of how problematic this may be for replicability, we tried to list as meticulously as possible the decisions that were taken and the rationale behind each one of them. Note also that information on risk intensity is lost, together with the effect of specific stressors, that of temporal trajectories and the possibility of statistical interactions between stressors. Furthermore, the inclusion of stressors in the score was inevitably limited by the availability of data, and, despite our best efforts to be as comprehensive as possible, it may still leave out important factors. Finally, the information included in the score relies predominately on mother reports, and although other information sources (e.g., partner or teacher reports) were included, when possible, reporter bias is still one of the key issues of this measurement.

Other applications
This score was originally conceived and used by Cecil et al., 2014; Rijlaarsdam et al., 2016; and Schuurmans et al., (in preparation). Similar versions of the score including prenatal and postnatal components have been succesfully used to predict later mismatch between cognitive abilities (i.e. IQ) and accademic achievement in Generation R children (Schuurmans et al., in preparation). Exposure to prenatal ELS was examined in relation to DNA methylation (DNAm) at birth, as a proxy of stable epigenetic changes (Rijlaarsdam et al., 2016), but no robust assocaitions we found. However, we are further examining whether prenatal ELS in interaction with genetic predisposition may have better explainatory power over DNAm patterns (Mulder et al., in preparation). Note that a parallel version of the postnatal ELS score, spanning from birth to 6 years of age, was created in Generation R and was examined in relation to internalizing and externalizing behavior (moderated by temperament and executive functioning; de Maat et al, in preparation) and problem behavior (measured with the Berkeley puppet interview).

Supplement 2 | Psycho-cardiometabolic comorbidity

This report describes the rationale and characteristics of the composite “comorbidity” outcome. The measure is intended as a sub-clinical proxy of comorbid depression and obesity in adolescence, and it captures the co-occurrence between high internalizing symptomatology (here referred to as “internalizing problems”) and high adiposity (i.e., “excess adiposity”). The scripts used to generate the measure in Generation R and ALSPAC are available here *[GitHub-link]*.

Internalizing symptoms score and adiposity at age 13 were weakly but significantly correlated in both cohorts. Pearson coefficients pooled across imputations were $r = 0.15$ (95%CI: [0.12; 0.18], $P = <.001$) in Generation R, and r = 0.11 (95%CI: [0.09; 0.14], $P = <.001$) in ALSPAC. Interestingly, the magnitude of these associations did not vary when android fat mass was considered as a proxy of adiposity in Generation R ($r = 0.15$, [0.12; 0.18]), but they were slightly reduced in ALSPAC ($r = 0.08$, [0.06; 0.10]). While these results do provide some evidence of an underlying linear association between internalizing and adiposity, it is hard to establish whether the

magnitude of this association is consistent with previous reports, because they predominantly rely on categorical classifications and less direct measures of adiposity (e.g., BMI)[12].

Hence, we computed psycho-cardiometabolic comorbidity as follows:
First, the two primary outcomes (i.e., internalizing symptoms and adiposity) were dichotomized into high and low-moderate, based on a sample-specific 80th percentile cut-off value. We decided to rely on a statistical cut-off value rather than a clinical one, to better capture subclinical patterns of comorbidity and to ensure comparability between different measures of internalizing symptoms available in the two cohorts (i.e., CBCL internalizing subscale in Generation R and SDQ emotional problems scale in ALSPAC).
The dichotomized values were then used to assign participants to one of four possible outcome groups:

1. "*Healthy*" if both internalizing symptoms and adiposity were below the 80th percentile value (marked with green in Figure 4 A and B). As expected, this was the largest group in both cohorts: pooled N across imputations was = 2791 (65% of the Generation R total sample) and 5916 (70.1% of ALSPAC total sample).
2. "*High internalizing*" if internalizing symptoms were above the 80th percentile but adiposity was low-moderate (marked with blue in Figure 4 A and B). Pooled N across imputations was = 623 (15% of the Generation R total sample) and 795 (9.4% of ALSPAC total sample).
3. "*High adiposity*" if adiposity was above the 80th percentile but internalizing was low-moderate (marked with yellow in Figure 4 A and B). Pooled N across imputations was = 631 (15% of the Generation R total sample) and 1476 (17.5% of ALSPAC total sample).
4. "*Comorbid*" if both outcomes were above the 80th percentile value (marked with red in Figure 4 A and B). This was the smallest group in both cohorts: pooled N across imputations was = 223 (5% of the Generation R total sample) and 241 (3% of ALSPAC total sample).

When comparing the relative group sizes in the two cohorts, one can appreciate how differences in both the univariate as well as bivariate distributions of the two outcomes were reflected in the categorical comorbidity outcome. In Generation R for instance, the healthy group was roughly 4 times larger than each of the individual outcome groups and 12 times larger than the comorbid group. Whereas in ALSPAC, a less symmetrical univariate distribution of the internalizing variable was reflected in the a "high internalizing" group that was about half the size of the "high adiposity" group. In addition, the decreased correlation between the two outcomes was reflected in a smaller comorbid group. Indeed, compared the healthy group, the high adiposity group was roughly 4 times smaller; the high internalizing group was 7 times smaller, and the comorbid group was 24 times smaller.

To assess whether the size of the comorbid group was larger than expected by chance alone, we performed a non-parametric permutation test. This procedure was described in detail elsewhere[13] and it was performed separately in each imputed separately, following these steps:

1) We first determined the size of the comorbid group (as described above).
2) We permuted the dichotomous "high internalizing" and "high adiposity" labels by randomly assigning participants to high (=1) or low-moderate (=0) internalizing and high (=1) or low-moderate (=0) adiposity. The original labels were shuffled such that the total number of high internalizing and high adiposity cases would not change, but their combination is now random.
3) We then combined the new random internalizing and adiposity labels to obtain a new "randomly comorbid" group, and we calculated its size.
4) We repeated step 2) and 3) 1000 times, thereby obtaining 1000 "randomly comorbid" group sizes.
5) We counted how often a number equal or larger than the original comorbid group size was observed under random permutation.
6) The resulting p-value (obtained by simply dividing the number of randomly equal or larger group sizes by 1000) reflected the probability that a comorbid group of the observed size would occur by chance alone[13].

In both cohorts, the permutation test confirmed that the size of the observed comorbid group was greater than expected by chance (pooled $P < .001$ in Generation R and $P = .006$ in ALSPAC).

Supplement 3 | Covariates description

This report describes the set of covariates included in the analyses. The scripts used to generate these measures in Generation R and ALSPAC are available here *[GitHub-link]*.

Child sex was measured at birth. The Generation R sample included 2087 males (48%) and 2181 (52%) females. The ALSPAC sample included 4370 males (52%) and 4058 females (48%).

1. **Child age** was measured at DXA visit (mean (years): 13.6 in Generation R and 13.8 in ALSPAC) and at completion of the CBCL or SDQ questionnaire (mean (years): 13.6 in Generation R and 13.2 in ALSPAC). A mean between the two age values was used as a covariate in the models, to account for: *(i)* its potential impact on both adiposity and internalizing symptoms; *(ii)* variable temporal gaps between ELS exposure (up to age 10 years) and outcome measurements (which could impact the strength of association).

2. **Child ethnic background**. In the Generation R cohort, ethnic background was determined by questionnaire-based assessment of the country of origin of participants' parents. Following Statistics Netherlands' guidelines[16], if one of the parents was born abroad, the child's ethnicity was determined according to that parent. If both parents were born abroad, the child was classified according to the mother's birthplace. Six large national groups were identified (i.e., Cape Verdean, Dutch, Dutch Antillean, Moroccan, Surinamese and Turkish). Smaller national groups were aggregated into five additional categories, according to the recommendations provided by the Joint Commitment for Action on Inclusion and Diversity in Scholarly Publishing (https://www.rsc.org/new-perspectives/talent/diversity-data-collection-in-scholarly-publishing/):
   - "*Africa and Middle East*" = Algeria, Angola, Armenia, Burundi, Cameroon, Congo, Côte d'Ivoire, Egypt, Eritrea, Ethiopia, French Congo, Gambia, Ghana, Guinea, Iran, Iraq, Israel, Kenya, Lebanon, Liberia, Mali, Mauritania, Mozambique, Nigeria, Palestine, Saudi Arabia, Senegal, Sierra Leone, Somalia, South Africa, Sudan, Syria, Tanzania, Togo, Tunisia, Uganda, Yemen, or Zimbabwe.
   - "*Asia and Oceania*" = Afghanistan, Australia, Bangladesh, China, Dutch New Guinea, East Timor, Hongkong, India, Indonesia, Japan, Kazakhstan, Korea, New Zealand, Pakistan, Philippines, Singapore, South Korea, Sri Lanka, Taiwan, Thailand, or Vietnam.
   - "*Europe*" = Austria, Belgium, Bosnia-Herzegovina, Canary Islands, Croatia, Czech Republic, Denmark, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Kosovo, Lithuania, Luxembourg, Madeira Islands, Moldova, Monaco, North Macedonia, Norway, Poland, Portugal, Romania, Russia, Serbia-Montenegro, Slovakia, Slovenia, Spain, Sweden, Switzerland, Ukraine, United Kingdom, or [ex-]Yugoslavia.
   - "*Latin America*" = Argentina, Belize, Bolivia, Brazil, Chile, Colombia, Cuba, Dominican Republic, Ecuador, Guyana, Haiti, Mexico, Paraguay, Peru, Trinidad and Tobago, or Venezuela.
   - "*North America*" = Canada or United States of America.

3. **Maternal pre-pregnancy BMI** was calculated from self-reported height and weight as weight / height^2.

4. **Maternal smoking during pregnancy** was self-reported and categorized into three levels:
   - never smoked,
   - smoked until pregnancy was known,
   - continued smoking during pregnancy.

5. **Maternal alcohol consumption during pregnancy** was measured differently in the two cohorts. The variable measured in Generation R comprised 4 levels:
   - never,
   - until pregnancy was known,
   - continued during pregnancy occasionally
   - continued during pregnancy frequently

   In ALSPAC alcohol consumption was measured twice during the first and last trimester of pregnancy. Each of these responses had six levels: 'never', '<1 glass per week', '1+ glasses per week', '1-2 glasses per day', '3-9 glasses per day', '10+ glasses per day'. We assigned a 0 to 5 severity score to each timepoint and then computed a mean across timepoints.

Supplement 4 | Imputation rationale and quality

Detailed strategy description

Missing values were imputed by fully conditional multiple imputation[17], using 60 iterations and 30 imputed datasets, as implemented by the mice package (version 3.13.0)[18]. The method of the imputation model was set to *predictive mean matching*, which is the recommended option in order to take into account non-normality (as well as sparse categorical data) [17,19], and it is relatively robust against model misspecification[17,20].

The auxiliary variables selected for the imputation model included: maternal age and smoking during pregnancy, gestational age and weight of the child at birth, parity, ethnicity and sex of the child, BMI of the mother during pregnancy and during childhood (measured when the child was 5), and maternal depressive symptoms during pregnancy and during childhood (measured when the child was 3).
Auxiliary variables were selected because they are believed to be either related to missingness or to the domain scores and primary outcomes themselves. When auxiliary information was available both prenatally and postnatally, the opposite period was used in the stress variables' imputation, in order to minimize bias. E.g., for the imputation of prenatal items, we used BMI of the mother when the child was 5, and for imputation of postnatal items we used BMI of the mother during pregnancy. This approach was meant to both reduce computational load and avoid multicollinearity issues.

Exposure imputation

Prenatal and postnatal stress scores were imputed in three steps: first the individual items were imputed according to the model specified below. Then mean domain scores were "passively" derived by averaging these complete indicators within their respective domain. At this stage a 25% cut-off was applied so that if less than 25% of the domain items was missing, only observed values were used, otherwise the imputed values were included in the calculation. Finally, the total prenatal and postnatal scores were calculated by summing the obtained domain scores within their respective periods.
The imputation of single ELS items was based on the following information:

(a) all other ELS items belonging to the same domain,
(b) other domain scores (excluding the domain of the imputed item),
(c) all auxiliary variables except for those that were part of the domain.

Covariates' imputation

Missing model covariates (age and ethnicity of the child, maternal BMI before pregnancy, maternal smoking and drinking behavior during pregnancy) were imputed given the outcome variables and the domain scores.

Outcome imputation

The primary outcomes (internalizing and adiposity at age 13) were imputed based on the total prenatal and postnatal stress scores, previous measurements of the same or correlated outcomes (assessed when the children were 10 years old), and the auxiliary variables.
Finally, the secondary "risk group" outcome (meat as a proxy of comorbidity) was imputed passively using the complete primary outcomes, as described in the main manuscript.

Exclusion criteria

We adopted a partial "imputation then deletion" approach for the exposure variables, whereas the imputed outcomes values were retained in the main analysis.
Participants were excluded from analysis when the frequency of missing ELS variables in the prenatal or the postnatal period exceeded 50%. All twins were further excluded. Finally, only one sibling was selected based on data completeness, or, if that was equal between the siblings, randomly.

Theoretical rationale

There are two important potential consequences of missing data. The first is the decrease in *precision* (wider confidence intervals) and *power* caused by the reduction in data. The second, and more serious, is the potential for *bias* in the estimation of association parameters[21-23].

Primer: missing data mechanisms

The complexity of the missing data problem, i.e., obtaining accurate inferential estimates in the presence of missing data, depends on the nature of the mechanism by which data are missing [21]. Little and Rubin [21,24] provided a popular framework to describe categories of missing data mechanisms given the relationship with observed and unobserved values.
The less problematic scenario occurs when the probability of an observable data point being missing does not depend on any observed or unobserved parameters: Missing Completely at Random (MCAR). More commonly, the missingness probability depends on observed variables, and hence it can be accounted for by the information

contained in the dataset: Missing at Random (MAR). The most challenging missingness mechanism occurs when the missingness probability depends on unobserved values: Missing Not at Random (MNAR).

Unfortunately, it is not possible to distinguish between MAR and MNAR mechanisms without additional external data or prior knowledge. However, the MAR assumption is usually reasonable in the context of longitudinal observational studies [23,25].
It is important to realize that the term *missing at random* does not mean that the missing data are a simple random subsample of all the data points. That scenario is MCAR. Under MAR, missing data may be more frequent in some subgroups than in others, but information defining the subgroups is observed [25].

Although missing data has the potential to cause serious bias, it is still possible to perform a valid and sensible analysis. Among the various available approaches to handling missing data[22,24,26], multiple imputation (MI) has been widely adopted and accepted by methodologists as an appropriate framework for dealing with MAR (and MCAR) mechanisms[17,21-23,25-27].

Primer: Multiple imputation
The multiple imputation approach adopted in this study can be described in three steps:
(1) Create multiple (M=30) copies of the dataset, with the missing values replaced by imputed values. To determine these values, the regression models described above are used to find cases with observed data that have predicted values closely resembling the predicted values of the respondents with missing values.
(2) Analyse each dataset separately with the chosen method (i.e., linear and multinomial logistic regression model, and causal mediation analysis).
(3) Pool the estimates and their standard errors across the M analyses using Rubin's rules[21], that allow to take into account the within-imputation and between-estimation variation components in the calculations. In this way, the uncertainty associated with imputation is accounted for.

The implicit and un-testable assumption is that the relationship between observed and missing values is the same for those who complete and those who do not: i.e., the MAR assumption.

A partial reason for MI's success is its flexibility[22]. MI is the only approach that can be used with any analytic model[17], which was particularly useful considered the range of variables and analytical approaches involved in this study.

Moreover, one of the most appealing features of the MI framework is the ability to incorporate additional "auxiliary" variables into the imputation model to improve the prediction of missing values. Incorporation of auxiliary data can make assumptions about the ignorability of the missing data more likely by reducing (or eliminating) bias[25,28]. The guiding principle in the construction of the above-specified models was indeed to leverage a wide range of information sources that could be predictive of the missing values themselves or influencing the process causing the missing data, even when these variables were not of interest in the substantive analysis.

For the imputation of exposure variables, the approach we have chosen involves two adaptations of the classical multiple imputation model: we defined a domain-specific set of predictors for individual item and we included a stepwise passive imputation for domain and total scores. This approach was recommended by van Buuren[17] as it has been found to reduce standard error substantially compared to complete-case analysis [29], and to outperform other existing techniques for handling large multi-item scales [30].

The outcome was included in the imputation model of missing covariate values, as recommended [25,31], however it was not included in the imputation of the stress predictors, although it might have improved the imputation performance [32]. This was done with the aim of creating a more generally valid stress score that could be used with a variety of outcomes in future studies, however this should be noted as a limitation.

As per the handling of missing outcome values, this was a particularly sensitive issue in this study, given the rather extended follow-up time that separated the baseline cohort inclusion from the outcome measurements (i.e., ~13 years). Unfortunately, discussion of these issues is lacking in the current literature and it is often unclear how missing data are being handled in practice in this context [22,33-35]. However, few simulation studies have shown that using information on exposure and correlated (e.g., longitudinally measured) outcomes in the imputation model reduced bias compared to alternatives[27,36-38] such as deleting cases with missing data or imputing and then deleting[37].

More broadly, although simpler solutions for handling missing outcomes are still routinely used[22,25,33-35], these approaches have been amply shown to be inadequate and even misleading[23,25-27,34,35], as they do not preserve important characteristics of the whole data set, such as key relationships among the variables and means[25,35]. For an overview of simulation studies that confirm that multiple imputation is a better alternative than listwise deletion and single imputation consult Van Buuren (2018)[17].

For instance, list-wise deletion (e.g., selecting on outcome availability prior to imputation) *requires MCAR data* in order to not introduce bias in the results[22,25], it makes strong assumptions about the covariance structure of the data (that it is compound symmetric), and it has been discouraged by statisticians [17,25]. In our study, we have reason to believe that dropout might depend at least partly on measured (or unmeasured) variables. For example, families that are exposed to lower levels of stress may be more likely to return follow-up questionnaires and bring their children to the visits. If that was indeed the case, the MCAR assumption would be violated and important concerns for selection bias would arise.
Deleting imputed outcomes prior to analysis can also lead to bias, especially when the imputation model contains variables that are associated with missingness in the outcome[37], as just described.

Hence, although it is important to point out that our approach relies on MAR assumptions, and we cannot guarantee unbiased results under the situation where missingness depends on unobserved information (MNAR), these assumptions are not nearly as strong (or in some cases, as unrealistic) as those required for a complete-case analysis. Moreover, it has been argued that MI can offer some protection against MNAR mechanisms [25,39], unlike MCAR methods [40], although this has not been quantified within a simulation framework. Nevertheless, MI can accommodate MNAR scenarios flexibly and is thus is well-suited to sensitivity analyses[41].

Missing patterns and imputation quality results
Table 4.1 provides some key variables' descriptives before (i.e., in the original sample) and after imputation (i.e., pooled across the 30 imputed datasets) together with the number and percentage of missing values in both cohorts.

Note that the pooled means (i.e., after imputation) of the exposure components and of the primary outcomes are slightly higher compared to the original metrics. We believe that this upward shift is to be expected under the assumption that the group lost to follow-up may be more likely to experience higher levels of stress, and higher physical and/or psychological problems. None of these differences resulted statistically significant.

The fraction of missing values per variable ranges from 0 to a maximum of 51% (i.e., the risk group variable in the ALSPAC cohort). Although the level of missingness naturally affects MI performance, we decided to leverage all available data even when missingness was extensive, as recommended by several statisticians [27,31,42].
In support of this approach, two simulation studies also suggested that, under the condition of a large enough sample size (<1000) and given that the imputation model is appropriate, such levels of missingness still allow for a reasonably low expected bias if any [27,38].

Convergence
Visual inspection of the convergence graphs showed no signs of unhealthy convergence. In some cases, the trace lines showed strong initial trends and slow mixing, but regardless of the proportion of missing data, results were stable after 20-30 of the 60 iterations. The convergence plots are not presented here, but are available upon request to SD.

Imputed vs. observed values
Next, we inspected and compared the density of the incomplete and imputed data. These graphs are publicly available for the key variables of interest and in the two cohorts (*[GitHub-link]*). The blue line marks the observed and red lines indicate imputed values.

Finally, to detect possible issues with the passive computation of the stress domain scores, we inspected the calculated domain values against imputed values These graphs are also publicly available for the key variables of interest and in the two cohorts (*[GitHub-link]*).

Sensitivity analysis

A thorough and sensible sensitivity analysis is an important step in producing and reporting robust estimates [22]. Hence, we also included a sensitivity analysis to assess the extent to which analytic approaches are robust to missing data assumptions. None of the main conclusions were impacted.

Table S1 | Assessment of missing values and imputation quality

See attached excel file (Supplementary Tables.xlsx).

Supplement 5 | Causal Mediation Analysis

Causal mediation analysis (CMA) is a method to dissect the total effect of an exposure into a direct and indirect effect (i.e., transmitted via the mediator to the outcome) [43,44]. We performed this analysis to disentangle the

underlying mediating role of postnatal stress in the relationship between prenatal stress exposure and each primary outcome (i.e., internalizing and adiposity respectively). The method was implemented using the CMAverse package [45] using a g-formula approach [46-48] to properly account for confounders of the mediator (i.e. postnatal ELS) – outcome relationship affected by the exposure (i.e. prenatal ELS). This consisted of the specification of two expectation statements. The first statement defines the outcome (i.e., internalizing and adiposity) as a function of prenatal ELS (i.e., the exposure), postnatal ELS (i.e., the mediator) and two sets of covariates:

- pre-exposure or baseline (i.e., sex, age, ethnicity, maternal BMI before pregnancy) and
- post-exposure or time-varying covariates (i.e., maternal smoking and drinking during pregnancy).

The second expectation defines the mediator, postnatal ELS, as a function of prenatal ELS and the two sets of covariates described above.

Because the exposure and mediator variable are continuous (i.e., prenatal and postnatal stress scores), reference levels for "no stress present" and "stress present" had to be specified. We decided to examine a change in prenatal stress from 0 (no stressor present) to 1. Hence, the active and control values for A, a and $a*$ correspond to prenatal ELS score = 0 (mean stress) and 1 (+1SD stress). Similarly, for postnatal stress, m is the value at which M is controlled (i.e., mean postnatal ELS). Consequently, G_a denotes a random draw from the distribution of M (i.e., postnatal ELS), had A=a (i.e., mean prenatal ELS), Y_{am} denotes the counterfactual value of Y that would have been observed had A been set to be a (i.e., mean prenatal ELS), and M to be m (i.e., mean postnatal ELS) and Y_{aGa*} denotes the counterfactual value of Y that would have been observed had A been set to be a, and M to be the counterfactual value G_{a*}.

With the g-formula approach, CMAverse estimates causal effects through direct counterfactual imputation estimation by the following steps:

1. For each post-exposure covariate q (i.e., maternal smoking and drinking during pregnancy), specify and fit a regression model for the distribution of L_q given A and C (e.g., maternal smoking ~ prenatal stress + ethnicity + maternal pre-pregnancy BMI)
2. For each post-exposure covariate q and for each individual i in the sample, simulate the counterfactuals $L_{a,q,i}$ and $L_{a*,q,i}$ from the regression models in step 1, by randomly drawing a value from the distribution of L_q given A=a (i.e., no prenatal ELS) or A=$a*$ (i.e., prenatal ELS) and C=C_i. Denote $L_{a,i}$=($L_{a,\text{smoking},i}$, $L_{a,\text{drinking},i}$)^T and $L_{a*,i}$=($L_{a*,\text{smoking},i}$, $L_{a*,\text{drinking},i}$)^T
3. Fit the regression model for the distribution of postnatal ELS (M) given A, L and C, i.e., postnatal ELS ~ prenatal ELS + covariates.
4. For each individual i, simulate the counterfactuals $M_{a,i}$ and $M_{a*,i}$ from the regression models in step 3, by randomly drawing a value from the distribution of M given A=a, L=$L_{a,i}$, C=C_i and given A=$a*$, L=$L_{a*,i}$, C=C_i.
5. Obtain $\{G_{a,i}\}$ by randomly permuting $\{M_{a,i}\}$ and $\{G_{a*,i}\}$ by randomly permuting $\{M_{a*,i}\}$.
6. Fit the regression model $E(Y \mid A,M,L,C)$ for each outcome Y (i.e., internalizing and adiposity).
7. For each individual i, using the regression model in step 6, obtain:
    - $E[Y_i \mid A=a*, M=m, L=L_{a*,i}, C=C_i]$,
    - $E[Y_i \mid A=a, M=m, L=L_{a,i}, C=C_i]$,
    - $E[Y_i \mid A=a*, M=G_{a*,i}, L=L_{a*,i}, C=C_i]$,
    - $E[Y_i \mid A=a*, M=G_{a,i}, L=L_{a*,i}, C=C_i]$,
    - $E[Y_i \mid A=a, M=G_{a*,i}, L=L_{a,i}, C=C_i]$ and
    - $E[Y_i \mid A=a, M=G_{a,i}, L=L_{a,i}, C=C_i]$.
8. These are then averaged across all participants to obtain the counterfactuals: $E[Y_{a*m}]$, $E[Y_{am}]$, $E[Y_{a*Ga*}]$, $E[Y_{aGa}]$, $E[Y_{aGa*}]$ and $E[Y_{a*Ga}]$, respectively.
9. Finally, the causal effects are calculated with the formulas reported in *Table 5.1*.

Because both outcomes were continuous, causal effects were estimated on the difference scale. Additionally, because of the time-varying confounders (L), natural direct and indirect effects are not identifiable, hence randomized interventional analogues of natural direct and indirect effects were used[47].

- Controlled Direct Effect (CDE) = $E[Y_{am}-Y_{a*m}]$;
- Randomized Analogue of Pure Natural Direct Effect (rPNDE) = $E[Y_{aGa*}-Y_{a*Ga*}]$;
- Randomized Analogue of Total Natural Direct Effect (rTNDE) = $E[Y_{aGa}-Y_{a*Ga}]$;
- Randomized Analogue of Pure Natural Indirect Effect (rPNIE) = $E[Y_{a*Ga}-Y_{a*Ga*}]$;
- Randomized Analogue of Total Natural Indirect Effect (rTNIE) = $E[Y_{aGa}-Y_{aGa*}]$;
- Total Effect (TE) = rPNDE+rTNIE or rTNDE+rPNIE

Confidence intervals were bootstrapped using 1000 repetitions.

This entire procedure was repeated in every imputed dataset and the estimates were pooled using Rubin's rules [49].

A sensitivity analysis was also conducted in the Generation R sample to assess the impact of potential unmeasured or unobserved confounders (U) that were associated with both postnatal stress (M) and each outcome Y (i.e., internalizing and adiposity). We used the mediational analog to the E-value [50], indicating the the minimum strength of association that an unmeasured confounder would need to have with both postnatal stress and the outcome of interest, to fully explain the reported estimates.

Because no robust approaches to pool E-value estimates across imputed dataset were available, we conducted the sensitivity analysis in a randomly selected imputed dataset.

The E-value for TE of prenatal stress on internalizing was 1.88 (lower limit of the CI = 1.79). This means that, if an unmeasured confounder was associated with both postnatal stress (M) and internalizing (Y), and the approximate risk ratios (RR) were 1.88 or more, the observed TE wold be completely explained.

The E-value for TE of prenatal stress on adiposity was 1.49 (lower limit of the CI = 1.39). This means that, if an unmeasured confounder was associated with both postnatal stress (M) and adiposity (Y), and the approximate RR were 1. 49 or more, the observed TE wold be completely explained.

The E-values for TE, rNDE and rNIE (on the RR scale) and their lower confidence interval limits were:

- Internalizing: TE = 1.88 (1.79); rNDE = 1.44 (1.33); rNIE = 1.59 (1.52)
- Adiposity: TE = 1.49 (1.39); rNDE = 1.34 (1.21); rNIE = 1.24 (1.15)

Note how the E-values for adiposity are relatively low, suggesting how these effects are in fact sensitive to unmeasured or unobserved confounders of the mediator-outcome association. The TE and rNIE on internalizing symptoms resulted more robust against such confounding.

Table S2 | Prenatal stress contribution (Causal Mediation Analysis results)

See attached excel file (Supplementary Tables.xlsx).

Table S3 | Postnatal stress contribution (Internalizing)

See attached excel file (Supplementary Tables.xlsx).

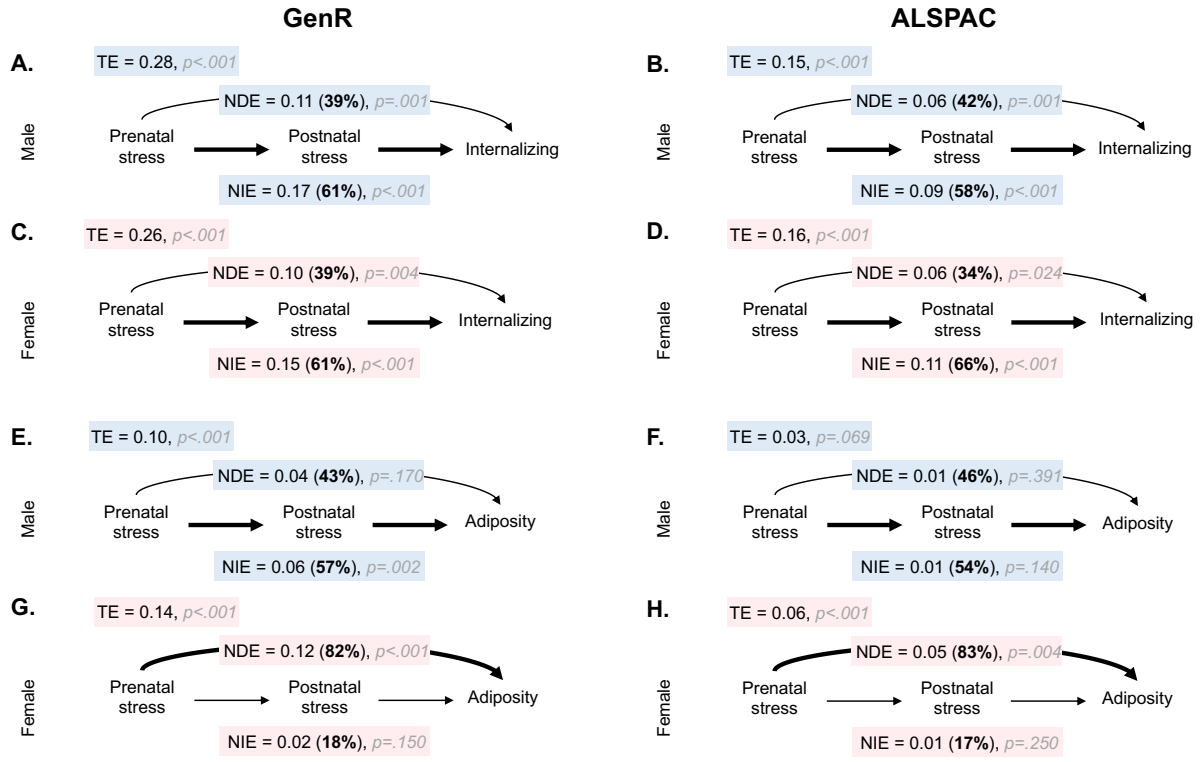
Table S4 | Postnatal stress contribution (Adiposity)

See attached excel file (Supplementary Tables.xlsx).

Table S5 | Comorbidity Analysis

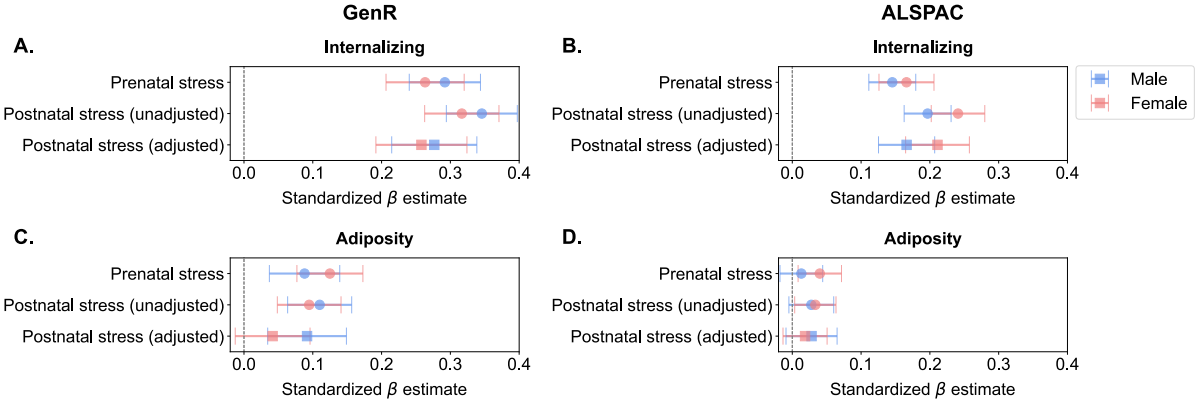
See attached excel file (Supplementary Tables.xlsx).

Figure S2 | Prenatal stress contribution (CMA) – stratified by sex



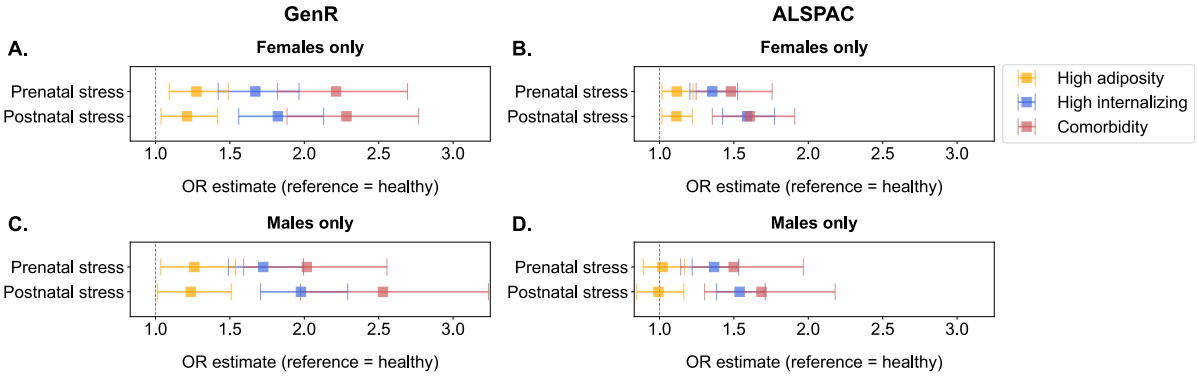
The causal estimates for the total effect (TE), natural direct (NDE) and natural indirect effect (NIE) of prenatal stress on internalizing (**A, C**: Generation R and **B, D**: ALSPAC) and adiposity (**E, G**: Generation R and **F, H**: ALSPAC) are displayed in the coloured boxes, with blue indicating male and pink indicating female strata. The percentage of the total effect due to direct and indirect pathway is reported between brackets and the predominant path is highlighted with a thicker arrow. P-values are reported in grey.

Figure S3 | Prenatal stress contribution (CMA) – stratified by sex



The *beta* estimates for pre- and postnatal ELS (and their 95% confidence intervals) are represented along the x-axis, and coloured according to the sex stratum (blue = males and pink = females). These represent the associations with internalizing symptoms in Generation R **(A)** and ALSPAC **(B)**, and adiposity in Generation R **(C)** and ALSPAC **(D)**.

Figure S4 | Comorbidity analysis – stratified by sex



The *beta* estimates for pre- and postnatal ELS (and their 95% confidence intervals) on the odds ratio (OR) scale are represented along the x-axis, with different colors depending on the comparison they refer to (yellow = healthy vs. high fat mass %; blue = healthy vs. high internalizing; red = healthy vs. comorbid), in Generation R males **(A)** and females **(C)** and ALSPAC males **(B)** and females **(D)**.

Table S6 | Prenatal stress contribution (Causal Mediation Analysis) - sex-stratified

See attached excel file (Supplementary Tables.xlsx).

Table S7 | Postnatal stress contribution (Internalizing) - sex-stratified

See attached excel file (Supplementary Tables.xlsx).

Table S8 | Postnatal stress contribution (Adiposity) - sex-stratified

See attached excel file (Supplementary Tables.xlsx).

Table S9 | Comorbidity analysis - sex-stratified

See attached excel file (Supplementary Tables.xlsx).

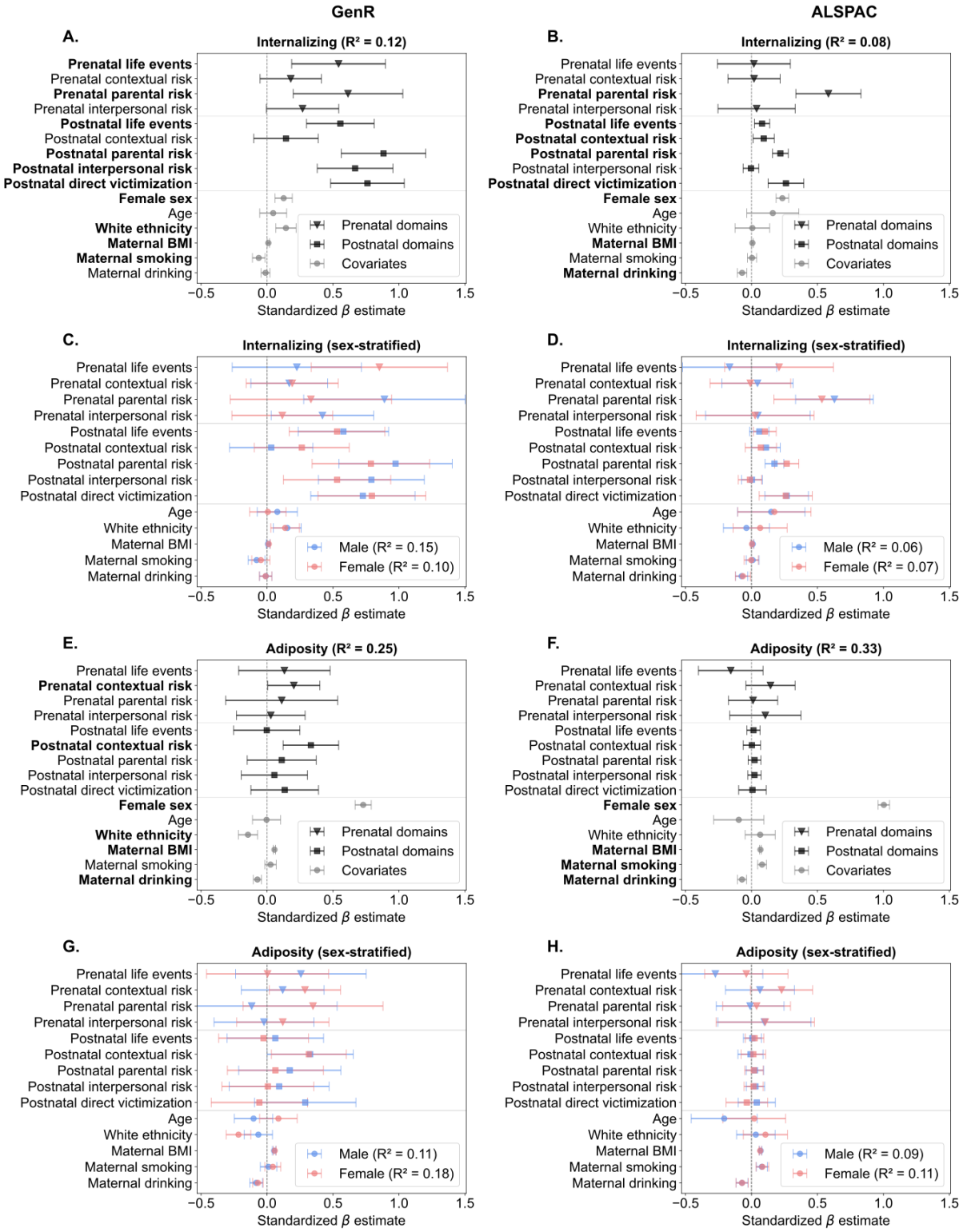
Figure S5 | Follow-up analysis: ethnic background * ELS interaction (GenR cohort)

(A) Pie chart of ethnic background groups identified in Generation R (with respective group size and percentage of the total sample).
(B) A stacked histogram representing the distribution of comorbidity status, i.e., percentage of children classified as heathy, or suffering from internalizing problems only, high adiposity only or comorbidity, in each ethnic group. Analogously, the two boxplot graphs depict the distribution of prenatal stress and postnatal stress respectively, across ethnic groups.
(C) Results of the follow-up analyses investigating the interaction between pre- and postnatal stress and ethnic background. The beta / OR estimates for prenatal (in grey) and postnatal ELS (in black) and their 95% confidence intervals are represented along the x-axis, separately for each ethnic background stratum. These represent the associations with internalizing symptoms, adiposity and their comorbidity. Note that the models assessing comorbidity as an outcome could not be estimated in the "North America" substratum due to insufficient sample size.

Table S10 | Follow-up analysis: ethnic background * ELS interaction (GenR cohort)

See attached excel file (Supplementary Tables.xlsx).

Figure S6 | Domain contribution analysis (internalizing and adiposity)

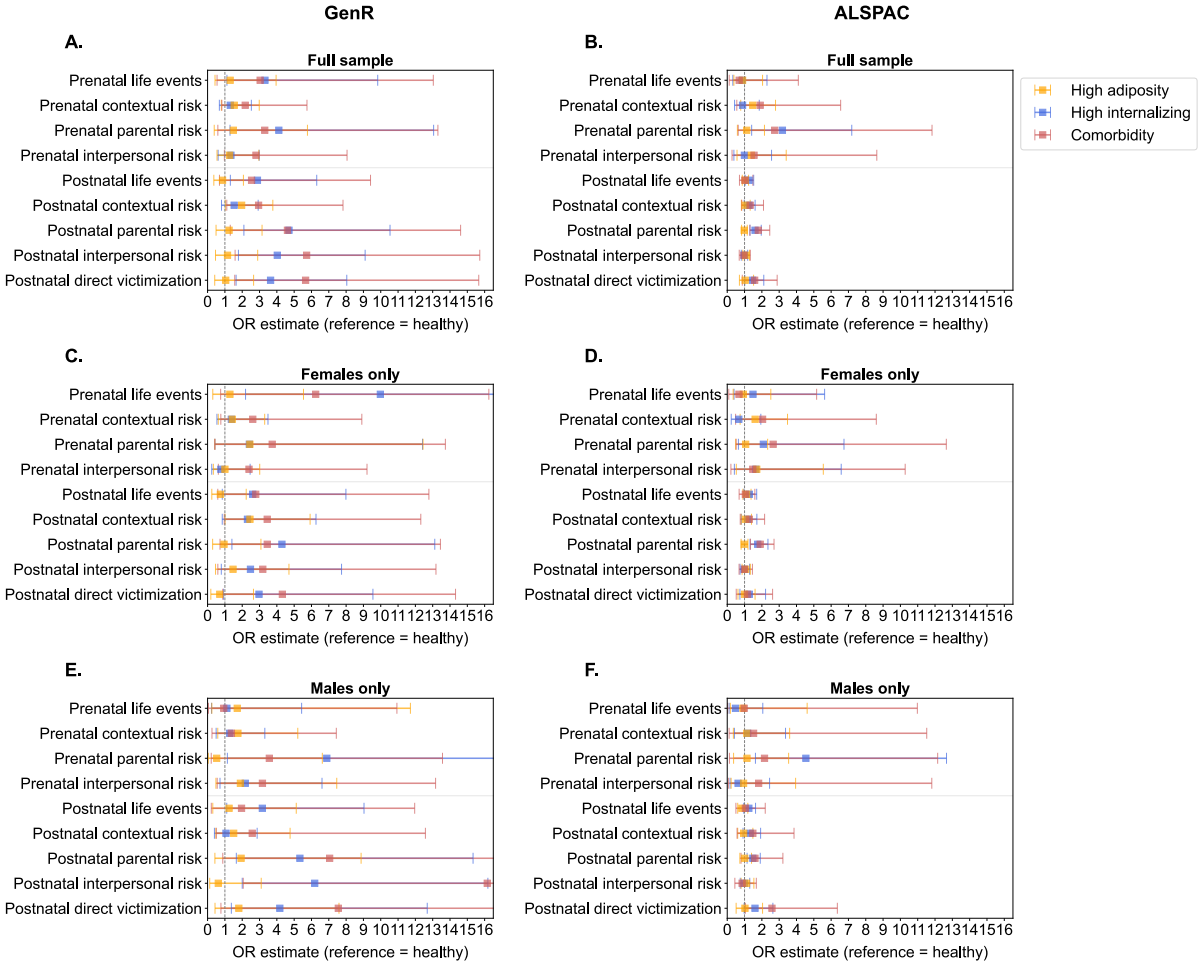


The *beta* estimates for each pre- and postnatal ELS domain (and their 95% confidence intervals) are represented along the x-axis for the prediction of internalizing in Generation R **(A)** and ALSPAC **(B)** full and sex-stratified samples **(C,D)**; and for that of adiposity in Generation R **(E)** and ALSPAC **(F)** full and sex-stratified samples **(G,H)**.

Table S11 | Domain contribution analysis (internalizing and adiposity)

See attached excel file (Supplementary Tables.xlsx).

Figure S7 | Domain contribution analysis (comorbidity)

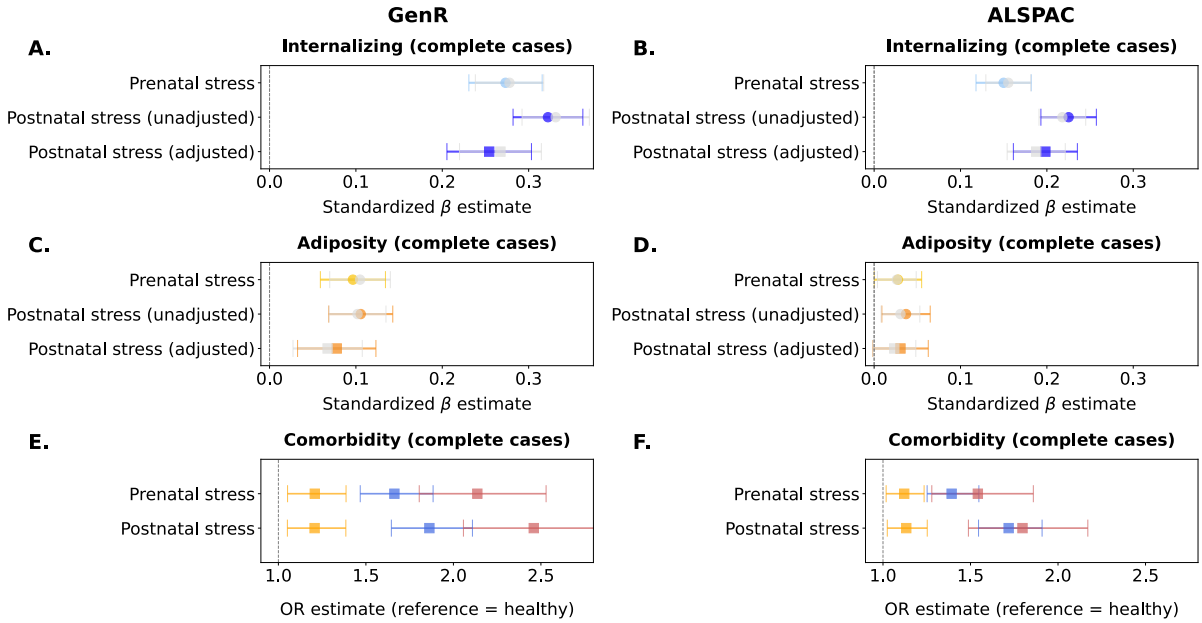


Effect estimates for each pre- and postnatal ELS domain (and their 95% confidence intervals) on the odds ratio (OR) scale are represented along the x-axis, with different colors depending on the comparison they refer to (yellow = healthy vs. high adiposity; blue = healthy vs. high internalizing; red = healthy vs. comorbid), in Generation R full sample **(A)**, females **(C)** and males only **(E)** and ALSPAC full sample **(B)**, females **(D)** and males only **(F).**

Table S12 | Domain contribution analysis (comorbidity)

See attached excel file (Supplementary Tables.xlsx).

Figure S8 | respondents–only (i.e., complete cases) sensitivity analysis



(A–D) The *beta* estimates for pre- and postnatal ELS (and their 95% confidence intervals) are represented along the x-axis for internalizing in Generation R **(A)** and ALSPAC **(B)** and adiposity in Generation R **(C)** and ALSPAC **(D),** *respondent-only samples*. These are complete cases, for which both outcomes were measured. Results of the main analyses are also displayed in grey for comparison.
(E, F) Effect estimates for pre- and postnatal stress (and their 95%CIs) on the odds ratio (OR) scale are represented along the x-axis, with different colours depending on the comparison they refer to (yellow = healthy vs. high adiposity; blue = healthy vs. high internalizing; red = healthy vs. comorbid), in Generation R **(E)** and ALSPAC **(F),** *respondent-only samples*.

Table S13 | Sensitivity analysis: respondents–only (internalizing)

See attached excel file (Supplementary Tables.xlsx).

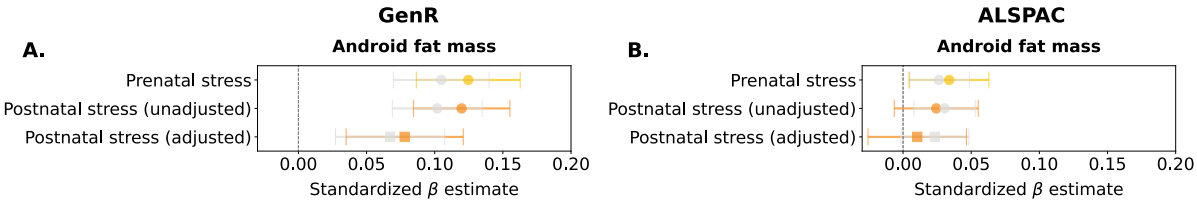
Table S14 | Sensitivity analysis: respondents–only (adiposity)

See attached excel file (Supplementary Tables.xlsx).

Table S15 | Sensitivity analysis: respondents–only (comorbidity)

See attached excel file (Supplementary Tables.xlsx).

Figure S8 | Android fat mass sensitivity analysis



The beta estimates for pre- and postnatal ELS (and their 95% confidence intervals) on android fat mass are represented along the x-axis in Generation R **(A)** and ALSPAC **(B)**. Results of the main analyses (i.e., using fat mass percentage as a measure of adiposity) are also displayed in grey for comparison.

Table S16 | Sensitivity analysis: android fat mass

See attached excel file (Supplementary Tables.xlsx).

Supplementary references

1. Cecil CA, Lysenko LJ, Jaffee SR, et al. Environmental risk, Oxytocin Receptor Gene (OXTR) methylation and youth callous-unemotional traits: a 13-year longitudinal study. *Molecular psychiatry*. Oct 2014;19(10):1071-7. doi:10.1038/mp.2014.95
2. Rijlaarsdam J, Pappa I, Walton E, et al. An epigenome-wide association meta-analysis of prenatal maternal stress in neonates: A model approach for replication. *Epigenetics*. 2016;11(2):140-9. doi:10.1080/15592294.2016.1145329
3. Heim C. Stress, Early Life. In: Gellman MD, Turner JR, eds. *Encyclopedia of Behavioral Medicine*. Springer New York; 2013:1903-1906.
4. De Beurs E. BSI: Brief symptom inventory. *Handleiding Addendum*. 2009;
5. Cox JL, Holden J, Henshaw C. *Perinatal mental health: the Edinburgh postnatal depression scale (EPDS) manual*. RCPsych publications London:; 2014.
6. Birtchnell J, Evans C, Kennard J. The total score of the Crown-Crisp Experiential Index: A useful and valid measure of psychoneurotic pathology. *British Journal of Medical Psychology*. 1988;61(3):255-266.
7. Byles J, Byrne C, Boyle MH, Offord DR. Ontario Child Health Study: Reliability and Validity of the General Functioning Subscale of the McMaster Family Assessment Device. *Family Process*. 1988;27(1):97-104. doi:https://doi.org/10.1111/j.1545-5300.1988.00097.x
8. Cortes Hidalgo AP, Neumann A, Bakermans-Kranenburg MJ, et al. Prenatal Maternal Stress and Child IQ. *Child Dev*. Mar 2020;91(2):347-365. doi:10.1111/cdev.13177
9. Muetzel RL, Mulder RH, Lamballais S, et al. Frequent Bullying Involvement and Brain Morphology in Children. Original Research. *Frontiers in Psychiatry*. 2019-September-24 2019;10doi:10.3389/fpsyt.2019.00696
10. Jansen PW, Raat H, Mackenbach JP, et al. Early Determinants of Maternal and Paternal Harsh Discipline: The Generation R Study. *Family Relations*. 2012;61(2):253-270. doi:https://doi.org/10.1111/j.1741-3729.2011.00691.x
11. Evans GW, Li D, Whipple SS. Cumulative risk and child development. *Psychol Bull*. Nov 2013;139(6):1342-96. doi:10.1037/a0031808
12. Halfon N, Larson K, Slusser W. Associations Between Obesity and Comorbid Mental Health, Developmental, and Physical Health Conditions in a Nationally Representative Sample of US Children Aged 10 to 17. *Academic Pediatrics*. 2013/01/01/ 2013;13(1):6-13. doi:https://doi.org/10.1016/j.acap.2012.10.007
13. Basso D, Pesarin F, Salmaso L, Solari A. Permutation Tests. *Permutation Tests for Stochastic Ordering and ANOVA: Theory and Applications with R*. Springer New York; 2009:1-35.
14. Mersky J, Topitzes J, Reynolds A. Impacts of adverse childhood experiences on health, mental health, and substance use in early adulthood: A cohort study of an urban, minority sample in the US. *Child abuse & neglect*. 2013;37(11):917-925.
15. McLaughlin KA, Hilt LM, Nolen-Hoeksema S. Racial/ethnic differences in internalizing and externalizing symptoms in adolescents. *J Abnorm Child Psychol*. Oct 2007;35(5):801-16. doi:10.1007/s10802-007-9128-1
16. Harding S, Teyhan A, Maynard MJ, Cruickshank JK. Ethnic differences in overweight and obesity in early adolescence in the MRC DASH study: the role of adolescent and parental lifestyle. *International journal of epidemiology*. 2008;37(1):162-172.

17. Van Buuren S. *Flexible imputation of missing data*. CRC press; 2018.
18. van Buuren S, Groothuis-Oudshoorn K. mice: Multivariate Imputation by Chained Equations in R. *Journal of Statistical Software*. 12/12 2011;45(3):1 - 67. doi:10.18637/jss.v045.i03
19. Marshall A, Altman DG, Royston P, Holder RL. Comparison of techniques for handling missing covariate data within prognostic modelling studies: a simulation study. *BMC Medical Research Methodology*. 2010/01/19 2010;10(1):7. doi:10.1186/1471-2288-10-7
20. Van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. *Journal of statistical computation and simulation*. 2006;76(12):1049-1064.
21. Rubin DB. Multiple imputation for survey nonresponse. New York: Wiley; 1987.
22. Bell ML, Fairclough DL. Practical and statistical issues in missing data for longitudinal patient-reported outcomes. *Stat Methods Med Res*. Oct 2014;23(5):440-59. doi:10.1177/0962280213476378
23. Sterne JA, White IR, Carlin JB, et al. Multiple imputation for missing data in epidemiological and clinical research: potential and pitfalls. *BMJ (Clinical research ed)*. 2009;338
24. Little RJ, Rubin DB. *Statistical analysis with missing data*. vol 793. John Wiley & Sons; 2019.
25. van Ginkel JR, Linting M, Rippe RCA, van der Voort A. Rebutting Existing Misconceptions About Multiple Imputation as a Method for Handling Missing Data. *Journal of Personality Assessment*. 2020/05/03 2020;102(3):297-308. doi:10.1080/00223891.2018.1530680
26. Graham JW, Cumsille PE, Shevock AE. Methods for handling missing data. *Handbook of psychology: Research methods in psychology, Vol 2, 2nd ed*. John Wiley & Sons, Inc.; 2013:109-141.
27. Kontopantelis E, White IR, Sperrin M, Buchan I. Outcome-sensitive multiple imputation: a simulation study. *BMC Medical Research Methodology*. 2017/01/09 2017;17(1):2. doi:10.1186/s12874-016-0281-5
28. Wang C, Hall CB. Correction of bias from non-random missing longitudinal data using auxiliary information. *Statistics in Medicine*. 2010;29(6):671-679. doi:https://doi.org/10.1002/sim.3821
29. Plumpton CO, Morris T, Hughes DA, White IR. Multiple imputation of multiple multi-item scales when a full imputation model is infeasible. *BMC Res Notes*. 2016;9:45-45. doi:10.1186/s13104-016-1853-5
30. Eekhout I, de Vet HC, de Boer MR, Twisk JW, Heymans MW. Passive imputation and parcel summaries are both valid to handle missing items in studies with many multi-item scales. *Statistical methods in medical research*. 2018;27(4):1128-1140.
31. Janssen KJ, Donders ART, Harrell Jr FE, et al. Missing covariate data in medical research: to impute is better than to ignore. *Journal of clinical epidemiology*. 2010;63(7):721-727.
32. Moons KG, Donders RA, Stijnen T, Harrell Jr FE. Using the outcome for imputation of missing predictor values was preferred. *Journal of clinical epidemiology*. 2006;59(10):1092-1101.
33. Wood AM, White IR, Thompson SG. Are missing outcome data adequately handled? A review of published randomized controlled trials in major medical journals. *Clin Trials*. 2004;1(4):368-76. doi:10.1191/1740774504cn032oa
34. Peugh JL, Enders CK. Missing Data in Educational Research: A Review of Reporting Practices and Suggestions for Improvement. *Review of Educational Research*. 2004;74(4):525-556. doi:10.3102/00346543074004525
35. Jelicić H, Phelps E, Lerner RM. Use of missing data methods in longitudinal studies: the persistence of bad practices in developmental psychology. *Dev Psychol*. Jul 2009;45(4):1195-9. doi:10.1037/a0015665
36. Floden L, Bell ML. Imputation strategies when a continuous outcome is to be dichotomized for responder analysis: a simulation study. *BMC Medical Research Methodology*. 2019/07/23 2019;19(1):161. doi:10.1186/s12874-019-0793-x
37. Sullivan TR, Salter AB, Ryan P, Lee KJ. Bias and Precision of the "Multiple Imputation, Then Deletion" Method for Dealing With Missing Outcome Data. *American Journal of Epidemiology*. 2015;182(6):528-534. doi:10.1093/aje/kwv100
38. Demirtas H, Freels SA, Yucel RM. Plausibility of multivariate normality assumption when multiply imputing non-Gaussian continuous outcomes: a simulation assessment. *Journal of Statistical Computation and Simulation*. 2008/02/01 2008;78(1):69-84. doi:10.1080/10629360600903866
39. Schafer JL, Graham JW. Missing data: our view of the state of the art. *Psychological methods*. 2002;7(2):147.
40. Molenberghs G, Thijs H, Jansen I, et al. Analyzing incomplete longitudinal clinical trial data. *Biostatistics*. Jul 2004;5(3):445-64. doi:10.1093/biostatistics/5.3.445
41. Groenwold RH, Donders ART, Roes KC, Harrell Jr FE, Moons KG. Dealing with missing outcome data in randomized trials and observational studies. *American journal of epidemiology*. 2012;175(3):210-217.
42. Madley-Dowd P, Hughes R, Tilling K, Heron J. The proportion of missing data should not be used to guide decisions on multiple imputation. *Journal of Clinical Epidemiology*. 2019/06/01/ 2019;110:63-73. doi:https://doi.org/10.1016/j.jclinepi.2019.02.016
43. Imai K, Keele L, Yamamoto T. Identification, inference and sensitivity analysis for causal mediation effects. *Statistical science*. 2010;25(1):51-71.

44. VanderWeele TJ, Vansteelandt S. Mediation Analysis with Multiple Mediators. *Epidemiol Methods*. 2014;2(1):95-115. doi:10.1515/em-2012-0010
45. Shi B, Choirat C, Coull BA, VanderWeele TJ, Valeri L. CMAverse: A Suite of Functions for Reproducible Causal Mediation Analyses. *Epidemiology*. 2021;32(5)
46. Robins J. A new approach to causal inference in mortality studies with a sustained exposure period—application to control of the healthy worker survivor effect. *Mathematical Modelling*. 1986/01/01/ 1986;7(9):1393-1512. doi:https://doi.org/10.1016/0270-0255(86)90088-6
47. VanderWeele TJ, Tchetgen EJ. Mediation analysis with time varying exposures and mediators. *J R Stat Soc Series B Stat Methodol*. Jun 2017;79(3):917-938. doi:10.1111/rssb.12194
48. Wang A, Arah OA. G-computation demonstration in causal mediation analysis. *Eur J Epidemiol*. Oct 2015;30(10):1119-27. doi:10.1007/s10654-015-0100-z
49. Rubin DB. *Multiple imputation for nonresponse in surveys*. vol 81. John Wiley & Sons; 2004.
50. VanderWeele TJ, Ding P. Sensitivity Analysis in Observational Research: Introducing the E-Value. *Ann Intern Med*. Aug 15 2017;167(4):268-274. doi:10.7326/m16-2607