Statistical challenges in identifying effects of age and age differences

Christian Dudel, D. Susie Lee, Angela Carollo, Nis Brix, Maria C. Magnus, Mikko Myrskylä

Abstract: The age difference between two persons is often of interest as an exposure in epidemiologic research. The association of the age difference between mother and father and child health outcomes is an example. In this paper, we show that there are three issues when using age differences as exposure. First, it is not possible to simultaneously control for the underlying age of both persons and the age difference without introducing arbitrary and untestable assumptions; instead, only two of these three variables can be accounted for at the same time. Second, we show formally that the age difference does not capture any interactions between the underlying ages; this might seem counterintuitive, given that the age difference is calculated using the age of both persons. Third, we argue that age differences might lack any deeper substantive meaning, at least in some applications. We illustrate these points using U.S. birth register data on more than 3 million births and modelling the influence of the maternal age, the paternal age, and the parental age difference on the risk of low birth weight; using this example, we provide recommendations on how to properly model and represent the interaction between age variables.

Key words: age, age difference, identification, maternal age, paternal age, birth weight

Introduction

The age difference between two persons is often of interest as an exposure in epidemiologic research. Examples include the association between the age difference between mother and father and child health outcomes (Yu et al. 2024; Wang et al. 2024); the age difference between partners and outcomes such as fecundability (Zhang et al. 2024), mortality (Syse & Lyngstad 2017; Drefahl 2010), or intimate partner violence (Adebowale 2018; Jewkes et al. 2006); the age difference between siblings and its role in child development (Havron et al. 2019; Eriksen et al. 2010); the age difference between organ donor and recipient and how it relates to the success of the transplantation (Ulrich et al. 2024); or the age difference between interviewer and survey participant and how it might alter response patterns (Okamoto et al. 2002). Moreover, there is work in which age differences are not the main exposure, but are included in the analysis as potential confounders; e.g., to control for partners' characteristics (Barclay et al. 2020; Copas et al. 2009). Interest in age differences also extends to fields beyond epidemiology such as sociology and biology (e.g., McKenzie 2021; Kuna et al. 2018; Tidière et al. 2018).

Intuitively, age differences are an appealing concept. They appear regularly in our everyday life. For instance, age differences between partners are a topic of conversation, and there are norms surrounding them (McKenzie 2021). Age differences between celebrities might garner considerable media attention, especially when they are perceived to deviate from these norms (Niccolai & Swauger 2021). Generally, age differences between partners vary between countries and cultures, and they have been changing over time (Ausubel et al. 2022; Dudel et al. 2021). Asking how such age differences are associated with health and other outcomes follows as an obvious question and is sometimes provided as a key motivation for studying age differences (e.g., Yu et al. 2024).

There are, however, three major challenges when studying age differences. Similar issues have been recognized in other contexts, such as the analysis of age-period-cohort effects (Osmond & Gardner 1989; Kupper et al. 1989), or when studying a change in a variable in addition to the values of the very same variable at two points in time (Preston et al. 2013). However, they have received little attention in the literature studying, or accounting for, age differences.

The first challenge is of statistical nature: there is a perfect collinearity between the two ages and the age difference. This is because the age difference is a function of the underlying ages. It is defined as the age of one person minus the age of another person; e.g., the age of the mother minus the age of the father. This makes it impossible to disentangle the individual association of these three variables with an outcome; or, technically speaking, only the effects of two of these three variables can be identified at the same time. Collinearity has received cursory mentions in the literature on age differences, but its consequences are usually not spelled out.

Second, as we show in this paper, the age difference does not capture the interaction between the underlying ages. This might be counterintuitive, as the age difference is calculated from two other variables which might make it seem like an interaction term. Instead, interactions between ages need to be explicitly modelled by multiplicative terms or similar means. Moreover, in a standard regression setting, including the age difference in addition to one of the underlying ages will always yield the same goodness-of-fit as controlling for both ages instead, while coefficient sizes and statistical significance will vary. This makes it difficult to justify the choice of variables solely from a statistical perspective, while at the same time the interpretation of results might differ depending on the choice.

The third challenge derives from the first and the second: what is the substantive meaning of the age difference given that it is fully dependent on the underlying ages and does not capture their interaction? Given the two statistical issues described above, the choice of variables can only be decided on a substantive basis. As we argue in this paper, in some cases it is difficult to interpret findings on age differences without recurring to the underlying variables, making the age difference of little relevance. We also briefly discuss cases in which age differences can be of substantive interest themselves. Moreover, we argue that age and age differences are, ultimately, often proxies for other key variables, and that it is preferable to measure these key variables directly whenever possible.

In this paper, we describe and discuss these three challenges in detail, and we provide a fully reproducible empirical example. It studies the maternal age, the paternal age, and the parental age difference as exposures and the risk of a low birth weight as the outcome. There is well-established evidence that both the maternal age and the paternal age impact child outcomes through several biological and social channels (Goisis et al. 2018; Andersen & Urhoj 2017), and recent research has started to study the relevance of age differences for child health outcomes at birth (Yu et al. 2024; Wang et al. 2024). Our example uses U.S. vital registration data for 2023 and it illustrates how analyses accounting for two ages and their interaction can be conducted, and that it is straightforward to infer the association of age differences with the outcome from such analyses if needed.

The identification problem: take two, get one free

Why is there perfect collinearity and why does it lead to statistical issues? As an example, we will consider the parental age difference calculated as the maternal age minus the paternal age, where both ages are measured in years. What follows applies to any age difference between two persons, where the ordering of persons can be either way, and it is only relevant for the sign of the effect of the age difference. In the example, the age difference could also be calculated as paternal age minus maternal age, and all results below would still hold.

Perfect collinearity immediately follows from the definition of the age difference. Formally, let D denote the parental age difference, M the age of the mother, and P the paternal age. They define each other through these equations:

$$D = M - P$$
$$M = D + P$$
$$P = M - D$$

These equations show that once two of these variables are known, the third can be perfectly predicted.

Perfect collinearity implies that it is not possible to identify the effects of all three variables on an outcome at once in a standard regression model. The issue is similar to the well-known identification problem for age-period-cohort models (Kupper et al. 1989). Specifically, the identification problem means that the regression coefficients of a model controlling for age, period, and cohort (or ages and the age difference) are not uniquely identified: different sets of regression coefficients, with potentially drastically different meaning, are equivalent with respect to standard objective functions, such as in OLS or maximum likelihood, and with respect to many measures of goodness-of-fit. Thus, it is not possible to determine a unique set of optimal values for the coefficients.

In case of the example of parental age and the parental age difference, the identification problem can be shown as follows. Let β_M, β_P, and β_D denote coefficient estimates for the regression

$$\hat{Y} = \alpha + \beta_M M + \beta_P P + \beta_D D,$$

where $\hat{Y}$ is the predicted value, and α is the constant. This model can be reparametrized as

$$\gamma_M = \beta_M + \lambda$$
$$\gamma_P = \beta_P - \lambda$$
$$\gamma_D = \beta_D - \lambda$$

where λ is an arbitrary constant; replacing β_M, β_P, and β_D with γ_M, γ_P, and γ_D yields

$$\begin{aligned}\hat{Y} &= \alpha + \gamma_M M + \gamma_P P + \gamma_D D \\ &= \alpha + \beta_M M + \beta_P P + \beta_D D + \lambda(M - P - D) \\ &= \alpha + \beta_M M + \beta_P P + \beta_D D.\end{aligned}$$

Thus, the two sets of coefficients lead to the very same predicted values, implying that the model is not uniquely identified, as standard objective functions will not differ.

For instance, assume that $\alpha = 0$, $\beta_M = 1$, $\beta_P = 1$, and $\beta_D = 0.5$; e.g., if the maternal age increases by one year, the outcome increases by one unit. Setting $M = 20$ and $P = 25$, which implies $D = -5$, the predicted value is $\hat{Y} = 42.5$. Setting $\lambda = -20$, we get $\gamma_M = -19$, $\gamma_P = 21$, and $\gamma_D = 19.5$. The coefficient for the maternal age now has a drastically different meaning: if the maternal age increases by one year, the outcome decreases by 19 units. However, $\hat{Y}$ still equals 42.5.

Generally, given that λ can be set at any arbitrary value, this implies that there is an infinite number of solutions to this regression problem. This applies to any generalized linear model, such as logistic regression, and it applies to any set of additive non-linear functions of the predictors, not just the linear case shown here (e.g., Osmond & Gardner 1989).

The identification problem cannot be solved or sidestepped. In case of the similar problem for age, period, and cohort, many solutions have been proposed. However, these all rely on more or less arbitrary, and often highly complex or hidden, assumptions (Bell 2020). These arbitrary assumptions have a big impact on the results (Robertson et al. 1999). Thus, no real solution exists.

Figure 1 provides a graphical representation of the identification problem and the collinearity of the maternal age, the paternal age, and the parental age difference. For simplicity, we assume that there only three ages 1, 2, and 3 for both mothers and fathers, where age 1 might represent young individuals, 2 middle-aged individuals, and 3 older individuals. All pairwise combinations of maternal age, paternal age, and the age difference are shown. Specifically, the left panel shows the maternal age on the x-axis and the paternal age on the y-axis, while the middle panel uses the maternal age on the x-axis and the age difference on the y-axis; the right panel shows the paternal age on the x-axis and again the age difference on the y-axis. In all panels, each cell represents one combination of ages. These ages are also written in each cell, plus the age difference: (M, P, D); e.g., the cell (2,1,1) refers to the case where the mother is in age group 2, the father is in age group 1, and the age difference is equal to 1. The cells are shown in different colors to make them easier to identify.

As can be seen from Figure 1, all three panels contain the same cells, just arranged in different ways, and any pair of values (M, P), (M, D), or (P, D) allows to identify the full cell (M, P, D) in all panels. This indicates that a two-dimensional representation is sufficient, and that adding a third dimension would not contain any additional information, as it is already implied by the other two dimensions.

Age differences are not interaction terms

Given perfect collinearity and the identification problem, what variables should researchers include in their analysis? Using only one of the three variables is often unsatisfactory, and might induce bias. For instance, in an analysis of child outcomes and how they depend on the parents, the maternal age and the paternal age will likely both be relevant covariates; at the same time, they are likely highly correlated. Assuming that, in addition, the effect of both parental ages is non-negligible and has the same sign, only including one of the two will bias the coefficient for the included age away from zero.

Thus, two of the three variables should be included. Three different combinations are possible. Researchers using the age of one person plus the age difference, instead of both ages, sometimes argue that using the age differences captures both underlying variables and how they might simultaneously shape outcomes, at least implicitly (e.g., Wang et al. 2024; Sandin et al. 2015). For instance, in case of the parental age difference calculated as maternal age minus paternal age, a value of -25 means that the father is 25 years older than the mother, and that there is a high likelihood that the mother is comparatively younger while the father is comparatively older, both with implications for child health outcomes and implying that the age difference captures, or at least proxies, both underlying ages. However, we show here that this is not true in a statistical sense.

Continuing our example of parental age differences, in a linear model, controlling for the maternal age (M) plus either the paternal age (P) or the age difference (D) is equivalent. Specifically, between the two regressions

$$\hat{Y} = \alpha + \beta_M M + \beta_P P$$

and

$$\hat{Y} = \delta + \gamma_M M + \gamma_D D,$$

the following relations hold (De Stavola et al. 2006):

$$\gamma_M = \beta_M + \beta_P$$
$$\gamma_D = \beta_P$$

$\delta = \alpha,$

allowing to transition from the first to the second equation, and conversely

$$\beta_M = \gamma_M + \gamma_D$$
$$\beta_P = -\gamma_D.$$

This means that there is a straightforward re-parametrization from one equation to the other, and both models will perform exactly similar with respect to metrics like R^2 and other measures of goodness-of-fit. Thus, having the age difference as a control in addition to maternal age does not capture more (or less) than controlling for the maternal age and the paternal age. This implies that the age difference does not function like a multiplicative interaction effect between the maternal age and the paternal age. Interaction effects need to be explicitly modelled, and we provide suggestions on how to do this in our empirical example. Moreover, it is not possible to decide between these two models (M-P vs. M-D) based on goodness-of-fit.

Importantly, the coefficients for maternal age from the two models, β_M and γ_M, do not need to be equal, can be of different sign and magnitude, and can also have different standard errors and p-values. For instance, if $\beta_M = \beta_P$ then γ_M will be twice as large as β_M. This means that, depending on the data, the choice of model could lead to different conclusions. In the supplementary materials we provide simulation results which demonstrate this for statistical significance.

Similar results hold if we start from the paternal age as a covariate and either add maternal age or the parental age difference. Moreover, all of the above relations also hold for generalized linear models such as logistic regression; but they do not necessarily apply when M, P, and D are not modelled linearly. Specifically, for such models, there is not necessarily a 1-to-1 mapping between the pairwise combinations of variables. This can apply, for instance, to piecewise constant models, in which age groups and/or groups of age differences are captured by dummies; e.g., a dummy for M being in the range of 20 to 24 and P being in the range of 25 to 29, or M being in the range of 20 to 24 and D being in the range of -4 to 0. Even though there is no perfect translation between models in this case, in practice the different models often approximate each other quite well and have extremely similar goodness-of-fit, as we show in our example application.

Substantive meanings of age differences

The previous results show that all three pairwise combinations of the ages and the age difference are statistically equivalent. What perspective to choose? Generally, age is not a cause in itself, but other causes operate along age (Berzuini & Clayton 1994); for instance, age is highly correlated with many biological processes and social processes. Similarly, age differences might serve as proxies for other factors; for instance, the age difference between a pair of siblings is an indicator to what extent we can expect them to grow up in a similar environment with shared experiences (Hemminki et al. 2008); another example is power relations and imbalances between partners which might be proxied by the age difference (Adebowale 2018). The question then is what underlying factors we are interested in and to base the choice on these factors. Generally, directly measuring these underlying factors would be preferable compared to using age and/or age differences as proxies, but might be difficult, whereas measuring age is straightforward.

How does this apply to the example of parental age and child outcomes at birth? There is well-established evidence that both the maternal age and the paternal age impact child outcomes through several biological and social channels (Goisis et al. 2018). For instance, in case of an older mother, there is an increased risk of negative outcomes for the child due to biological reasons; in case of a young father, there might be an increased risk as the father might be less well economically established; these two factors might interact, perhaps the first makes the child more vulnerable to the second, or vice versa.

This evidence could be used to argue that studying the parental age difference might be useful to understand the interplay of maternal and paternal age. However, as shown above, the age difference does not capture the interaction. Moreover, this argument does not assign any deeper meaning to the age difference beyond the underlying maternal and paternal age and how they might interact. Thus, in this example, it might be preferable to use the maternal age and the paternal age, and not the age difference, as the former have a much more straightforward interpretation.

Example: low birth weight and parental age

Data and methods

As an example, we study the association between the maternal age and paternal age and the risk of a low birth weight in the United States in 2023. The example is fully reproducible and uses R (R Core Team 2025). Code is available on GitHub (https://github.com/christiandudel/age_difference), while the data is freely available from the National Bureau of Economic Research (https://www.nber.org/data/vital-statistics-natality-data.html).

The birth register contains information on all live births for which birth certificates have been filed in the U.S. during 2023. In our analysis, we only include births for which the mothers resided in the U.S.; for which both the maternal age and the paternal age are known (both measured in years); and for which the birth weight was recorded. This reduces the number of observations by around 10% and leaves us with 3,214,954 births out of the total of 3,605,081 births. We consider a child to have low birth weight if the weight was below 2,500 grams (ICD-11 codes KA21.0, KA21.1, and KA21.2Z; WHO 2025).

Using the U.S. birth register data, we fitted 9 logistic regression models. In all models, the outcome is a binary indicator of low birth weight. The first model accounts for the maternal age and the paternal age using only linear terms; the second model is non-linear and adds quadratic terms of maternal age and paternal age, as well as an interaction term of maternal age and paternal age, and a squared interaction; the third model is piecewise-constant and uses dummy variables for all combinations of 5-year age intervals, such as for mothers being 20 to 24 and fathers being 25 to 29. The fourth, fifth, and sixth model are similar to the first three, except that the paternal age is replaced with the parental age difference. Models seven to nine control for the paternal age and the parental age difference in a similar fashion. Results for models which include quadratic terms but no interactions are discussed in the supplementary materials. To visualize the results, we will use contour plots of the predicted probability of a low birth weight.

Results

Figure 2 shows a visualization of the raw data, similar to Figure 1. In the left panel, each cell represents a combination of maternal age and paternal age; for each age, the color indicates the unadjusted probability of a low birth weight, calculated only using the births with the exact combination of maternal and paternal age: dark blue indicates a low risk of a low birth weight, shades of green indicate a moderate risk, and yellow indicates a high risk. Age combinations which are represented with less than 30 observations are shown in grey, and age combinations which do not appear in the data are left blank. The middle panel can be read similarly, with the exception that the y-axis shows the parental age difference; and the right panel shows the parental age difference on the y-axis and the paternal age on the x-axis. In all three panels, outcomes for fathers aged 45 are shown as black lines, as an example how the three panels correspond to each other. In the left panel, the line is horizontal, in the middle panel it is diagonal, and in the right panel it is vertical.

As can be seen from Figure 2, there is an increased risk for a low birth weight when the age of the mother is young or old; when the age of the father is young or old; or when both conditions apply (mother young/old and father young/old). With respect to age differences, both comparatively low and high age differences go along with an increased risk; moderate age differences, such as around -5 years, show both low and high levels of risk.

Results of the logistic regression models are shown in Figure 3. Specifically, the predicted probability of a low birth weight is shown. All models are displayed on the same x-axis (maternal age) and y-axis (paternal age); for the models including the parental age difference, this means that the results were re-arranged. Unadjusted, raw results like in Figure 2 are shown as a reference at the bottom; as in Figure2, for the raw results cells with less than 30 observations are shown in grey. Predicted values are only shown for cells which are observed in the data.

All linear models perform poorly. The predicted probabilities do not vary much for two reasons. First, the relationship between the maternal age/paternal age and outcomes at childbirth is often nonlinear and U- or J-shaped (e.g., Urhoj et al. 2017), and this also applies in this case. Fitting a straight regression line to such a pattern leads to lines which are horizontal or almost horizontal with a low slope; as a result, the predicted outcome does only increase very slightly with increasing maternal or paternal age. Second, the models do not capture the interaction of maternal and paternal, missing additional nuance. Models which include nonlinear terms but no interaction perform better than purely linear models, but worse than the more complex models; for details, see the supplementary materials.

Both the nonlinear interacted models and the piecewise-constant models perform very well in capturing key properties of the data. This is also shown in model summary statistics such as the Akaike Information Criterion (AIC). In all cases, differences between using the paternal age or the age difference are non-existent (linear) or marginal (non-linear, piecewise constant). For instance, for the piecewise constant models, the AIC equals 1,781,567 (maternal age, paternal age), 1,781,657 (maternal age, age difference), and 1,781,740 (paternal age, age difference), respectively. For the linear models, the AIC is 1,785,619 in all cases and thus indicates a worse fit compared to the piecewise models.

Conclusions

Including age differences in a quantitative analysis should be considered carefully. Age differences do not capture the interaction of the underlying ages, and whether they are of substantive interest will depend on the application. In case of parental age differences and child outcomes we argue that age differences might provide a less useful perspective on the outcome than the maternal age and the paternal age. Irrespective of the choice of variables, explicitly modelling their interaction is crucial. Contour plots provide an effective way for representing and interpreting the results, as shown by our example and by earlier work (e.g., Yu et al. 2024; Bille et al. 2005). While this paper focuses on age and the age difference as the key exposure of interest, our recommendations also apply to cases in which they are used as additional control variables and in which they are not of a major concern themselves. Overall, our paper shows that researchers should take great care to consider how they handle age differences in their analysis.

References

Adebowale, A.S. (2018): Spousal age difference and associated predictors of intimate partner violence in Nigeria. *BMC Public Health* 18, 212. https://doi.org/10.1186/s12889-018-5118-1

Andersen, A.-M. N., Urhoj, S. K. (2017): Is advanced paternal age a health risk for the offspring? *Fertility and Sterility* 107, 312-318. https://doi.org/10.1016/j.fertnstert.2016.12.019

Ausubel, J., Kramer, S., Shi, A. F., & Hackett, C. (2022): Measuring age differences among different-sex couples: Across religions and 130 countries, men are older than their female partners. *Population Studies*, *76*, 465–476. https://doi.org/10.1080/00324728.2022.2094452

Barclay, K. J., Thorén, R. D., Hanson, H. A., Smith, K. R. (2020): The Effects of Marital Status, Fertility, and Bereavement on Adult Mortality in Polygamous and Monogamous Households: Evidence From the Utah Population Database. *Demography* 57: 2169-2198. https://doi.org/10.1007/s13524-020-00918-z

Bell, A. (2020): Age period cohort analysis: a review of what we should and shouldn't do. *Annals of Human Biology* 47, 208-217. https://doi.org/10.1080/03014460.2019.1707872

Berzuini, C., Clayton, D. (1994): Bayesian analysis of survival on multiple time scales. *Statistics in Medicine* 13, 823-838. https://doi.org/10.1002%2Fsim.4780130804

Bille, C., Skytthe, A., Vach, W., Knudsen, L. B., Andersen, A.-M. N., Murray, J. C., Christensen, K. (2005): Parent's Age and the Risk of Oral Clefts. *Epidemiology* 16, 311-316. https://doi.org/10.1097/01.ede.0000158745.84019.c2

Copas, A. J., Mercer, C. H., Farewell, V. T., Nanchahal, K., Johnson, A. M. (2009): Recent Heterosexual Partnerships and Patterns of Condom Use. A Weighted Analysis. *Epidemiology* 20, 44-51. https://doi.org/10.1097/EDE.0b013e318187ac81

De Stavola, B., Nitsch, D., dos Santos Silva, I., McCormack, V., Hardy, R., Mann, V., Cole, T. J., Morton, S., Leon, D. A. (2006): Statistical Issues in Life Course Epidemiology. *American Journal of Epidemiology* 163, 84-96. https://doi.org/10.1093/aje/kwj003

Drefahl, S. (2010): How does the age gap between partners affect their survival? *Demography* 47, 313-326. https://doi.org/10.1353/dem.0.0106

Dudel, C., Cheng, Y.-h.A. and Klüsener, S. (2023), Shifting Parental Age Differences in High-Income Countries: Insights and Implications. *Population and Development Review* 49, 879-908. https://doi.org/10.1111/padr.12597

Eriksen, W., Sundet, J. M., Tambs, K. (2010): Birth Weight Standardized to Gestational Age and Intelligence in Young Adulthood: A Register-based Birth Cohort Study of Male Siblings. *American Journal of Epidemiology* 172, 530–536. https://doi.org/10.1093/aje/kwq199

Goisis, A., Remes, H., Barclay, K., Martikainen, P., Myrskylä, M. (2018): Paternal age and the risk of low birth weight and preterm delivery: a Finnish register-based study. *Journal of Epidemiology and Community Health* 72, 1104-1109. https://doi.org/10.1136/jech-2017-210170

Havron, N., Ramus, F., Heude, B., Forhan, A., Cristia, A., Peyre, H., Annesi-Maesano, I., Bernard, J. Y., Botton, J., Charles, M. A., Dargent-Molina, P., de Lauzon-Guillain, B., Ducimetière, P., De Agostini, M., Foliguet, B., Forhan, A., Fritel, X., Germa, A., Thiebaugeorges, O. (2019): The Effect of Older Siblings on Language Development as a Function of Age Difference and Sex. *Psychological Science* 30, 1333-1343. https://doi.org/10.1177/0956797619861436

Hemminki, K., Li, X., Sundquist, K., Sundquist, J. (2008): Familial risks for chronic obstructive pulmonary disease among siblings based on hospitalisations in Sweden. *Journal of Epidemiology & Community Health* 62 : 398-401. https://doi.org/10.1136/jech.2007.063156

Jewkes, R., Dunkle, K., Nduna, M., Levin, J., Jama, N., Khuzwayo, N., Koss, M., Puren, A., Duvvury, N. (2006): Factors associated with HIV sero-status in young rural South African women: connections between intimate partner violence and HIV. *International Journal of Epidemiology* 35, 1461-1468. https://doi.org/10.1093/ije/dyl218

Kuna, B., Galbarczyk, A., Klimek, M., Nenko, I., Jasienska, G. (2018): Age difference between parents influences parity and number of sons. *American Journal of Human Biology* 30, e23095. https://doi.org/10.1002/ajhb.23095

Kupper, L. L., Janis, J. M., Karmous, A., Greenberg, B. G. (1985): Statistical age-period-cohort analysis: A review and critique. *Journal of Chronic Diseases* 38, 811-830. https://doi.org/10.1016/0021-9681(85)90105-5

McKenzie, L. (2021): Age-dissimilar couple relationships: 25 years in review. *Journal of Family Theory & Review* 13, 496-514. https://doi.org/10.1111/jftr.12427

Niccolai, A., Swauger, M. (2021): Minding the (Age) Gap: The Identity and Emotion Work of Men and Women in Age-Discrepant Romantic Relationships. *Sociological Focus 54*, 19-38. https://doi.org/10.1080/00380237.2020.1845258

Okamoto, K., Ohsuka, K., Shiraishi, T., Hukazawa, E., Wakasugi, S., Furuta, Kayoko (2002): Comparability of epidemiological information between self- and interviewer-administered questionnaires. *Journal of Clinical Epidemiology* 55, 505-511. https://doi.org/10.1016/S0895-4356(01)00515-7

Osmond, C., Gardner, M. J. (1989): Age, Period, and Cohort Models. Non-Overlapping Cohorts Do Not Resolve the Identification Problem. *American Journal of Epidemiology* 129, 31-35. https://doi.org/10.1093/oxfordjournals.aje.a115121

Preston, S. H., Mehta, N. K., Stokes, A. (2013): Modeling Obesity Histories in Cohort Analyses of Health and Mortality. *Epidemiology* 24, 158-166. https://doi.org/10.1097/EDE.0b013e3182770217

R Core Team (2025): R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

Robertson, C., Gandini, S., Boyle, P. (1999): Age-Period-Cohort Models: A Comparative Study of Available Methodologies. *Journal of Clinical Epidemiology* 52, 569-583. https://doi.org/10.1016/S0895-4356(99)00033-5

Sandin, S., Schendel, D., Magnusson, P., Hultman, C., Surén, P., Susser, E., Grønborg, T., Gissler, M., Gunnes, N., Gross, R., Henning, M., Bresnahan, M., Sourander, A., Hornig, M., Carter, K., Francis, R., Parner, E., Leonard, H., Rosanoff, M., Stoltenberg, C., Reichenberg, A. (2016): Autism risk associated with parental age and with increasing difference in age between the parents. *Molecular Psychiatry* 21: 693-700. https://doi.org/10.1038/mp.2015.70

Syse, A., Lyngstad, T. H. (2017): In sickness and in health: The role of marital partners in cancer survival. *Social Science & Medicine: Population Health* 3, 99-110. https://doi.org/10.1016/j.ssmph.2016.12.007

Tidière, M., Thevenot, X., Deligiannopoulou, A., Douay, G., Whipple, M., Siberchicot, A., Gaillard, J.-M., Lemaître, J. F. (2018): Maternal reproductive senescence shapes the fitness consequences of the parental age difference in ruffed lemurs. *Proceedings of the Royal Society* B 285, 20181479. https://doi.org/10.1098/rspb.2018.1479

Ulrich, S., Arnold, L., Michel, S., Tengler, A., Rosenthal, L., Hausleiter, J., Mueller, C. S., Schnabel, B., Stark, K., Rizas, K., Grabmaier, U., Mehilli, J., Jakob, A., Fischer, M., Birnbaum, J., Hagl, C., Massberg, S., Haas, N., Pozza, R. D., Orban, M. (2024): Influence of donor age and donor-recipient age difference on intimal hyperplasia in pediatric patients with young and adult donors vs. adult patients after heart transplantation. *Clinical Research in Cardiology*. https://doi.org/10.1007/s00392-024-02477-4

Urhoj, S. K., Andersen, P. K., Mortensen, L. H., Smith, D. G., Andersen, A.-M. N. (2017): Advanced paternal age and stillbirth rate: a nationwide register-based cohort study of 944,031 pregnancies in Denmark. *European Journal of Epidemiology* 32, 227-234. https://doi.org/10.1007/s10654-017-0237-z

Wang, S.-H., Lin, M.-C., Wu, C.-S., Chen, P.-C., Thompson, W. K., Fan, C.-C. (2024): Familial factors rather than paternal age contribute to the aetiology of epilepsy. *International Journal of Epidemiology* 53, dyad191. https://doi.org/10.1093/ije/dyad191

World Health Organization (2025): International Classification of Diseases for Mortality and Morbidity Statistics. Eleventh Revision. https://icd.who.int/, accessed Sept 16, 2025.

Yu, V. T., Ramsay, J. M., Horns, J. J., Mumford, S. L., Bruno, A. M., Hotaling, J. (2024): The association between parental age differences and perinatal outcomes. *Human Reproduction* 39, 425-435. https://doi.org/10.1093/humrep/dead236

Zhang, Y., Zhang, H., Zhao, J., Zhao, Y., Zhang, J., Jiang, L., Wang, Y., Peng, Z., Zhang, Y., Jiao, K., He, T., Wang, Q., Shen, H., Zhang, Y., Yan, D., Ma, X. (2024): Gravidity modifies the associations of age and spousal age difference with couple's fecundability: a large cohort study from China. *Human Reproduction* 39, 201-208. https://doi.org/10.1093/humrep/dead209

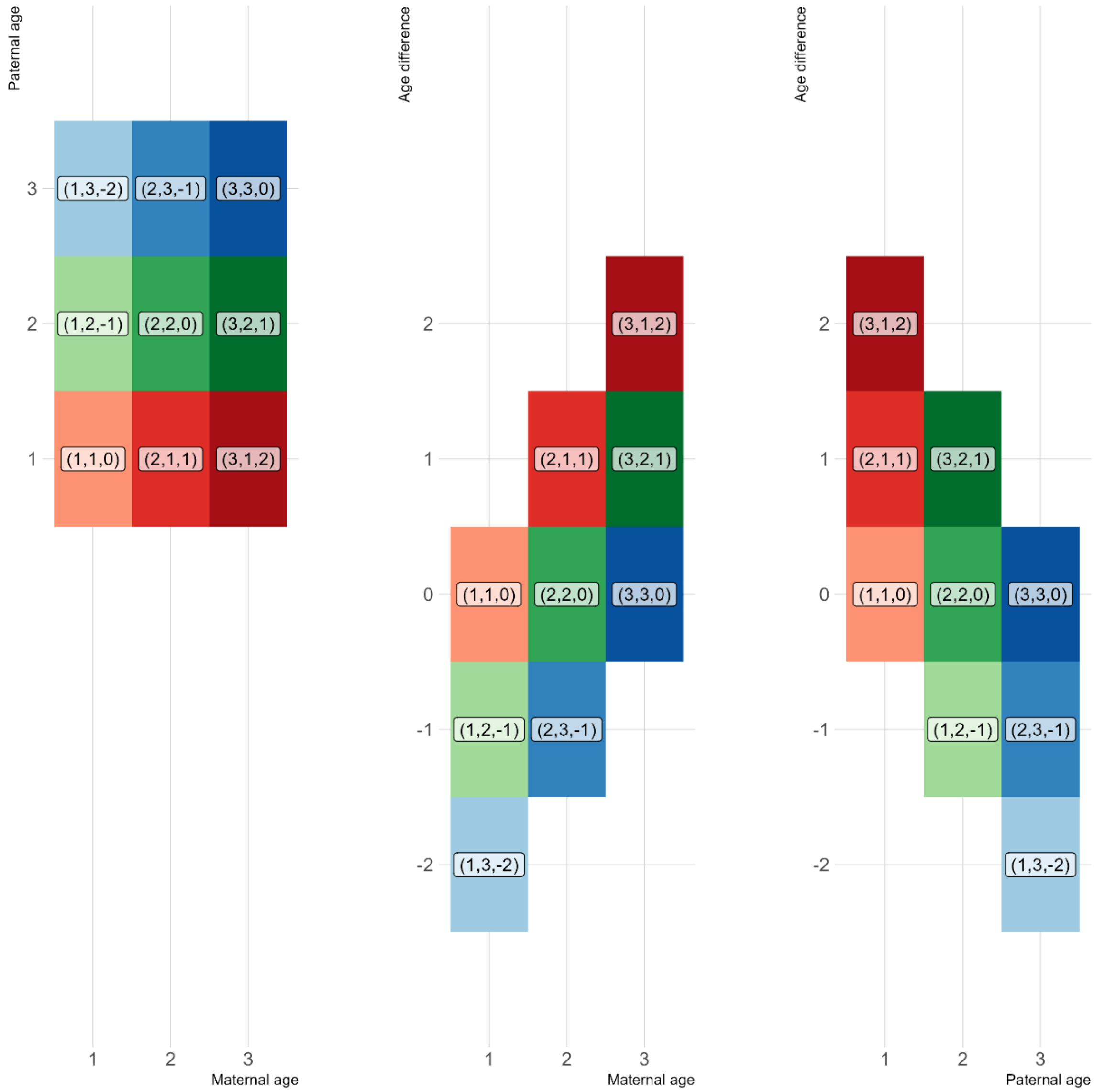

Figure 1: A graphical representation of the identification problem. To keep this example simple, there are three maternal and paternal ages (1, 2, 3). Each cell represents one combination of maternal (M) and paternal (P) age, as well as the implied age difference (D). They are shown in each cell as (M,P,D). Colors are used to distinguish cells. All pairwise combinations of maternal age, paternal age, and the age difference are shown. Specifically, the left panel shows the maternal age on the x-axis and the paternal age on the y-axis, while the middle panel uses the maternal age on the x-axis and the age difference on the y-axis; the right panel shows the paternal age on the x-axis and again the age difference on the y-axis.

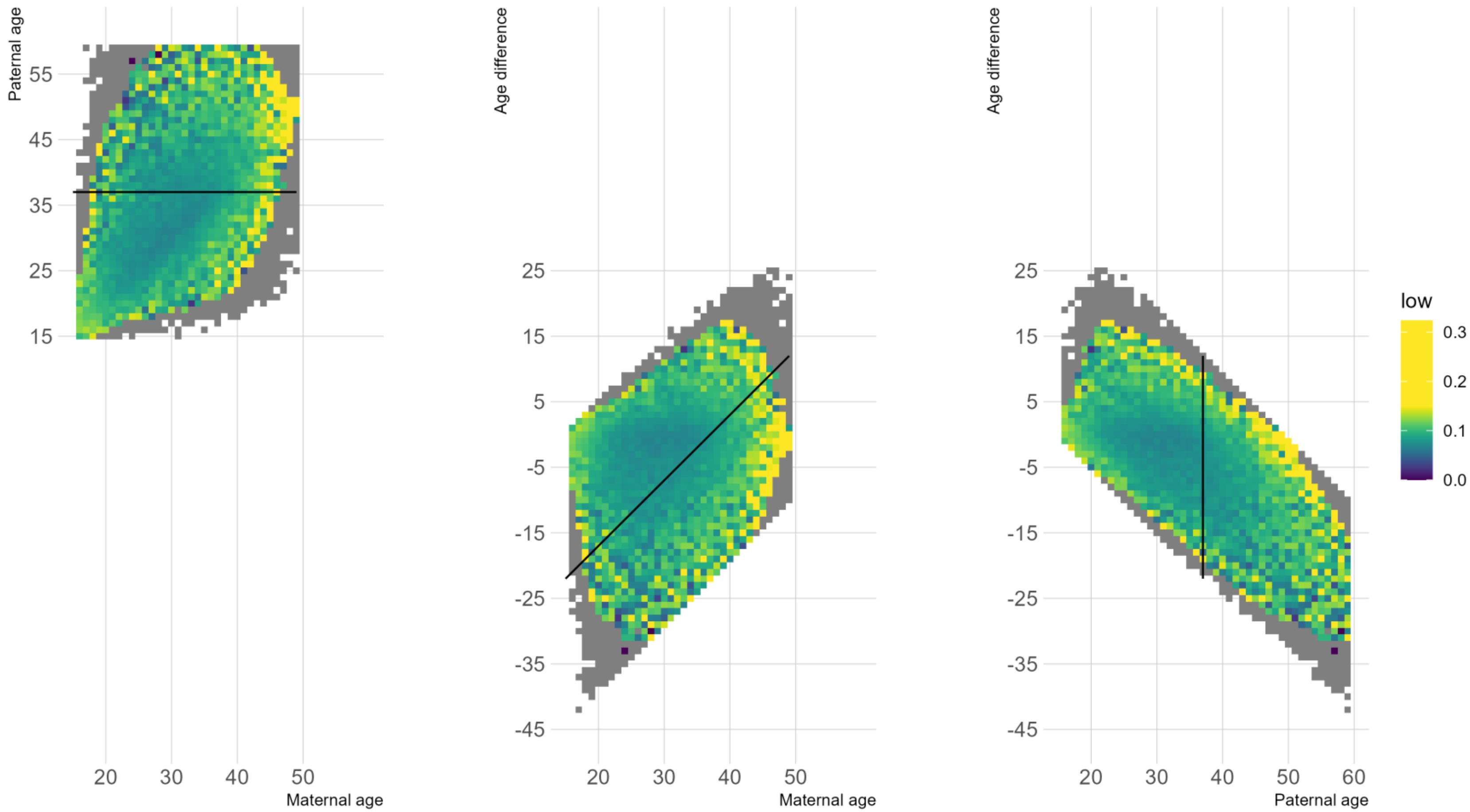

Figure 2: Unadjusted probability of a low birth weight (<2500g) by maternal age (M), paternal age (P), and the parental age difference (M-P). Cells with less than 30 observations shown in grey.

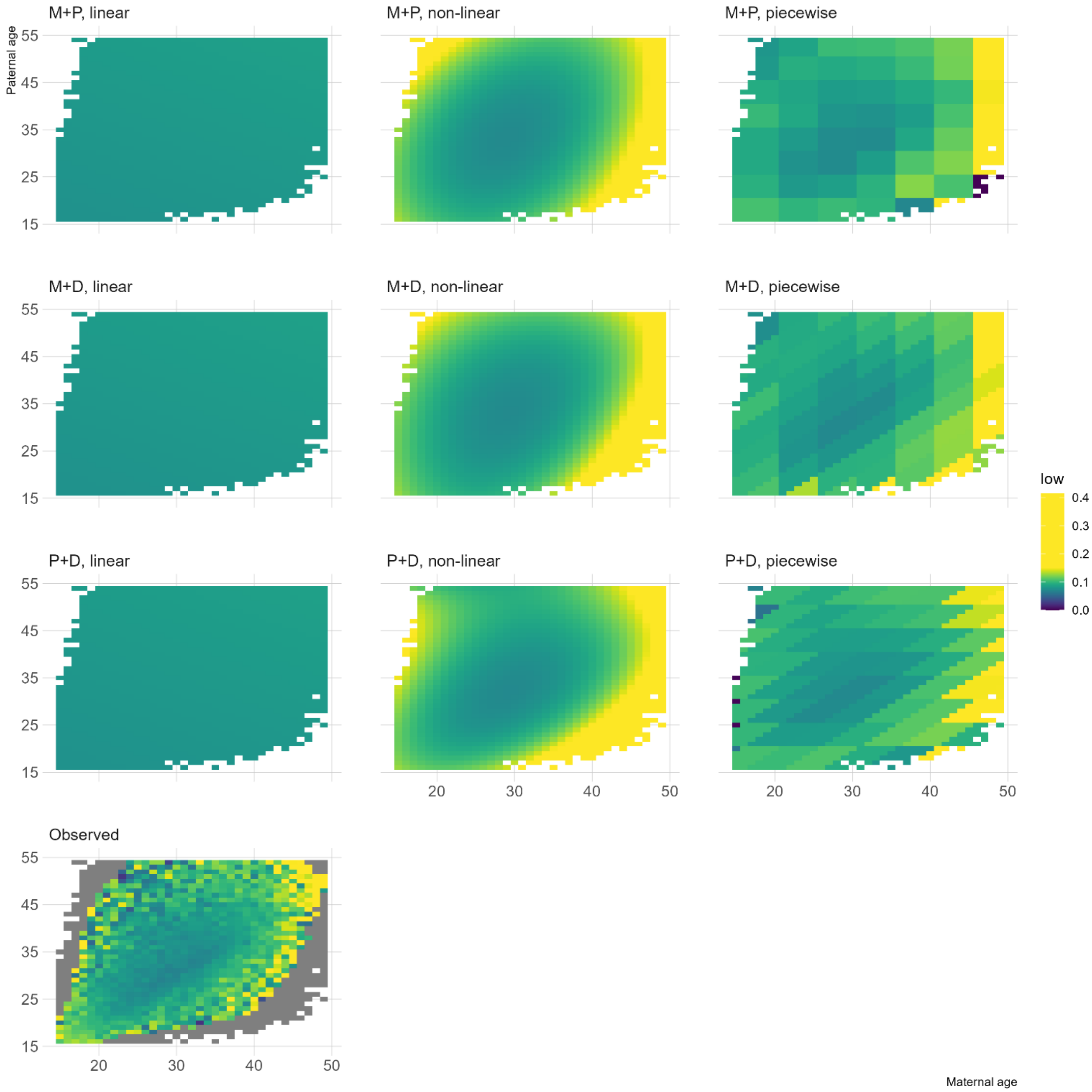

Figure 3: Predicted risk of low birth weight. "M+P" controlling for maternal and paternal age, "M+D" controlling for maternal age and parental age difference, and "P+D" controlling for paternal age and parental age difference.

Supplementary materials – simulation results

As discussed in the main text, the results of statistical inference can depend on the parametrization; for instance, the coefficient for the effect of the maternal age could be statistically significant at the 5% level when also controlling for the parental age difference, but not when controlling for the paternal age instead.

To show this, we generated data sets in which the data generating process for risk of a low birth weight (or any binary outcome) is derived from a linear probability model with

$$W_i = 0.07 + 0.0005M_i + 0.00075P_i + \epsilon_i,$$

where W_i is the risk of observation i to have low birth weight; M_i and P_i are the maternal and paternal age of observation i, respectively; and ϵ_i is a random term. Given a sample size N, we draw random and independent ages for the mother and the father from a uniform distribution on the integers of 20 to 45; and ϵ_i is drawn from a folded normal distribution with expectation zero and a standard deviation of 0.1. For each observation, we randomly set a binary variable to one – i.e., low birth weight – with probability W_i.

We create data sets for sample sizes of 250, 500, 750, …, 100,000; and for each sample size, we create a total of 500 data sets. On each data set, we calculate logistic regression models controlling either for M and P, or for M and D. For each sample size, we provide two sets of results: the average p-value of the coefficient for M over all 500 data sets of the same sample size; and the probability that the p-value for M is below the conventional 5% threshold in the model controlling for M and D but not in the model controlling for M and P, also over all 500 data sets.

Results are shown in Figure S1. In the upper panel, it can be seen that the p-value of the coefficient for M drops much faster below 5% (shown as a dashed grey line) in the model accounting for M and D compared to the model controlling for M and P. On average, the M-D model reaches a p-value below 5% around a sample size of 17,000 observations, while for the M-P model this only happens around 52,000 observations. The consequence of this is shown in the lower panel: assume a researcher would base their decision of the importance of maternal age mainly on whether the coefficient is statistically significant at the 5% level. The results show that given the simulation setup, for a large range of sample sizes, there is a probability of up to 50% that the two different models lead to different conclusions.

This result does depend on the coefficients for M and P, and it is possible to generate scenarios in which both types of models yield very similar results, or in which both models are even further apart. In an applied setting, a researcher will only know when explicitly comparing models. As shown in our example in the main text, this issue does not apply when using models properly accounting for interactions because these will usually lead to very similar results irrespective which combination of age and/or age difference is used.

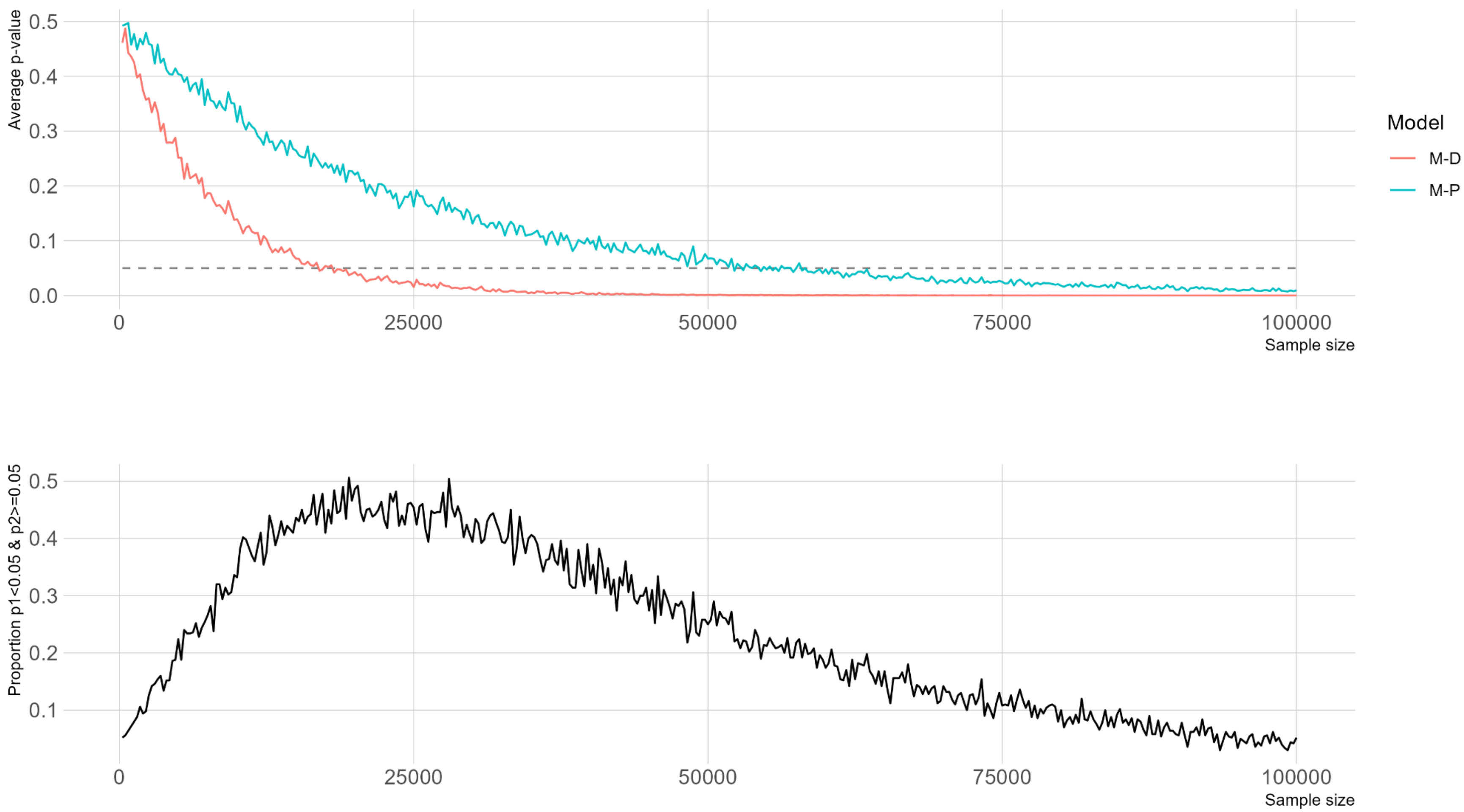

Figure S1: Simulation results. Upper panel shows p-value of the coefficient of M in the model controlling for maternal age and the age difference (red) and the model controlling for the maternal age and the paternal age (M-P; blue); 5% threshold shown as dashed horizontal line. The lower panel shows the proportion of simulations in which the coefficient for maternal age is significant for the M-D model but not for the M-P model.

Supplementary materials – non-linear models without interaction

Figure S2 shows result similar to Figure 3 in the main text. On the left side, it includes the results from the nonlinear models with interaction terms which are shown in the rightmost column in Figure 3. In the right column, it includes nonlinear models which include quadratic terms for both predictors, but no interaction. While the nonlinear, interacted models on the left all provide very similar results, the nonlinear models without interaction perform very differently. While the model including the maternal age and the age difference is not far off from the interacted model, this is not true for the other two variants. In particular, controlling for the paternal age and the age difference without interaction misses key features.

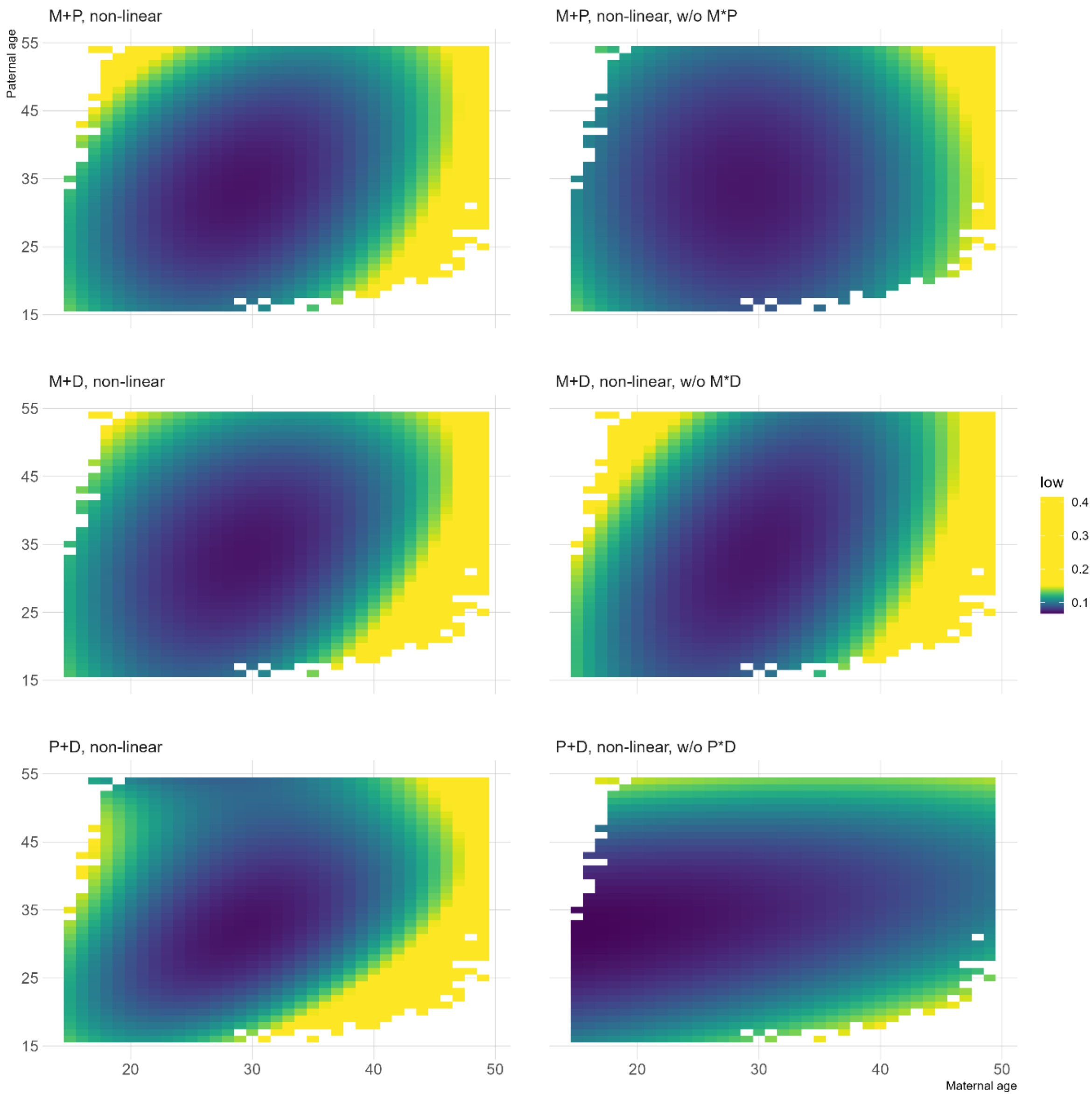

Figure S2: Prediced probability of low birth weight by maternal age and paternal age, based on nonlinear models with interaction (left) vs. nonlinear models without interaction (right).