

Epigenetic aging and lifespan reflect reproductive history in the Finnish Twin Cohort

Corresponding Author: Ms Mikaela Hukkanen

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

This paper provides a new type of evidence from a large data set of monozygotic and dizygotic twins that supports the evolutionary theory of aging in some detail. I am comfortable with the statistical analyses, but I would like to suggest a few changes to the citations:

This paper also provides genetic evidence, of a different kind, supporting the evolutionary theory of aging:

Long, Erping; Zhang, Jianzhi. Evidence for the Role of Selection for Reproductively Advantageous Alleles in Human Aging. *Obstetrical & Gynecological Survey* 79(3):p 165-166, March 2024. | DOI: 10.1097/01.ogx.0001010448.83325.67

Line 47: This is a better reference to the current state of evolutionary medicine (your reference 10):

Stearns, S.C. & Medzhitov, R. 2024. *Evolutionary Medicine* 2nd Edition. Oxford University Press. 400 p. ISBN-10: 0192871985

Line 54: This paper provides additional evidence of a U-shaped mortality curve:

Wang, X., Byars, S.G., & Stearns, S.C. 2013. Genetic links between post-reproductive lifespan and family size in Framingham. *Evolution, Medicine, and Public Health*. 2013. doi: 10.1093/emph/eot013

For future research: are at least some of the loci identified by Long and Zhang also identified in the epigenetic data sets?

Reviewer #2

(Remarks to the Author)

This article reports analysis of associations of reproductive history with survival and biological aging in a large Finnish Twin cohort. The main findings are that women who have had more pregnancies tend to be biologically older, as measured by the GrimAge epigenetic clock, and have shorter lifespans as compared with women who have had fewer pregnancies. In addition, they find that early childbearing or being nulliparous (never having children) are also associated with these phenotypes of accelerated aging.

This paper makes a contribution to the literature on reproductive history and biological aging. George Williams theorized that aging represented a tradeoff between reproduction and somatic maintenance. This theory was later refined by Thomas Kirkwood. Here, the researchers report evidence expanding on findings by Ryan et al. There is longstanding epidemiological evidence of shorter lifespans and higher morbidity among women with more children or more pregnancies. More recently, Ryan et al. along with others have reported evidence of accelerated biological aging as measured by epigenetic clocks. This paper adds to that literature by including an analysis of sibling differences that partly controls for genetic confounding of associations between reproductive history and biological aging.

I have a few comments that, being addressed, would enhance the impact of the article.

The authors quantify reproductive history using a latent class model to assign women to categories based on their probability of live birth in each of a series of 3-year time windows. The intuition for why this approach is superior to simply counting the number of live births or alternative specifications of fertility history is not clear from the text. The authors should offer more explanation for why their chosen method is superior to more straightforward approaches. They should also report results from simple tabular approaches somewhere in the article.

It's not clear why zygosity is not considered in the adjusted models that include family fixed effects (meaning the models that include dummy variables for each family ID). Perhaps analysis could be stratified by zygosity?

It's nice that the authors report the twin correlations for GrimAge, although this is a bit off topic from the paper. If they choose to keep this in the paper, they should report it for the other clocks as well.

Some sensitivity analysis testing whether results are consistent for women measured on different DNAm platforms would be useful.

Reviewer #3

(Remarks to the Author)

Overall: This manuscript reports on an analysis of data from Finland that aimed to examine latent classes of pregnancy history in women accounting for timing and number, and examine these statistically derived classes with mortality risk, and also examine a convenience sample of epigenetic based data in relation to the derived classes. Overall, there are many methodologic considerations that appear unaddressed that limit a clear interpretation of the results for the reader and limit a placing of this research in the context of what is actually a fairly robust evidence base (when considering myriad chronic disease conditions) related to the same exposure. Combined with inappropriate causal language interspersed throughout, and overexuberant interpretation, there is limited enthusiasm in the current form.

Specific Comments

Abstract

The language implies that a causal effect was estimated, but the nature of the question and the methods applied to answer it are not estimating an effect, thus "impact" is inappropriate language as are further conclusions beyond the sentence below this statement noting the objective is to examine an association.

What does "contemporary" supposed to mean to the reader? What about "early" or "numerous". This is vague/unclear language for the reader.

The conclusive statement does not reflect the research and its not clear why emotive language is included.

Main body-Intro

1st paragraph – causal language not supported by evidence – those are associations noted, not effects. Ironically, the last sentence of that paragraph undercuts/contradicts the earlier notions. Which is it? This is confusing for the reader. As noted, this isn't a new topic/idea and it has strong theory behind it, along with multiple studies supporting the theory.

Need some references to support statement that no study has examined age of first pregnancy and age/timing/count of any subsequent pregnancies. This statement also implies that there has been a comparison of approaches in the current study – do the data support this in the current study?

What is being communicated re: confounders? This meandering statement is unclear and could be interpreted multiple ways. Have previous studies not accounted for likely confounders?

Methods

What is the overall N of women in the study?

How many were sent this questionnaire in 1967? It appears 17,000+ responded – but its important to know for selection bias purposes, particularly since this wasn't clearly sampled and was a convenience response. Any conclusion is predicated on understanding the selection of the population under study and what is described would suggests a likely biased sample.

Its unclear how a cohort that includes women born in the dates noted could be considered "contemporary".

It also appears the questionnaire was sent to 9-10 year old twins in 1967. How can a latent class analysis be appropriately classified if using data where its possible a number of participants were not even biologically capable of being pregnant, and unlikely to be so for a number of years?

Given the different birth cohorts, it seems results presented by these would be merited to make sound conclusions.

It's unclear how well classified and valid the exposure is for the latest birth cohort, given they were close to 18 when the data started, as well as hadn't had a chance to have the same exposure/patterns as groups they were compared to or apart of the same latent class.

A major issue with interpreting the results is that the assessment of many factors is either predicated on survival, or likely meaning something completely different depending on the lifestage (for example, BMI in 1975 – could be 90+ years or 18...)

When was alcohol assessed?

Please report the N of overall and each variable included in multiple imputation. Related, was the data evaluated for the nature of missingness? This is possible and would inform the nature of the imputation, rather than just assuming the most minimal biased approach as reported.

How was the epigenetic data sampled- random? Convenience? Some sort of applied sampling approach? When?

Given the strong confounding effect of factors like age, different generations, and others, plus potential selection biases on who was sampled, how is the reader supposed to interpret the epigenetic data? Shouldn't the reader be interested in the comparison of people of very similar age ranges to make any sort of conclusion?

It appears the survival data was predicated on participating in the 1975 assessment – what is the N relative to the beginning sample?

Survival analysis

-What is the sample size and selection for this analysis? The truncation summary suggests that there was a need to participate in 1967 and be alive in 1975, which suggests that this was the start of follow up or time 0?

-The follow up time calculation suggests that no participant was lost to follow up – is this the case?

-For the modeling, for the youngest participants, the variable of lifetime education at the 1975 date is likely to misclassify this confounder.

-Why was C3 the reference group? There is no logic or rationale provided.

Results, etc..

Supplemental doesn't really suggest any higher risks, or higher epigenetic based age calculations except perhaps in c6 class, but also in nulliparous group. The main results, with imputation, are largely suggestive except for a signal in the nulliparous group. How does this support the over the top conclusions noted by the authors? This is misleading.

What is the reader supposed to make of the differential distribution of participant exposure classes by self-report, v. selected epigenetic based measures?

Residual confounding due to lack of more comprehensive measurements is a likely strong potential explanation for the findings.

Is the interpretation of the survival data as "longevity" appropriate? These were never compared w/in the birth cohorts, and probably can't reliably be compared except in the latest (cold war) cohort.

Given the unclear comparability considerations around the epigenetic based data, it's not clear what they add.

Overall, there is no balanced discussion of the science, methods, results and no mention of any limitations or other considerations.

Reviewer #4

(Remarks to the Author)

The study sets out to test whether there is an association between reproduction patterns (number and timing of children) and survival in 17,080 Finnish twin women and further in a subset of 1,117 twins to test whether there is an association between reproductive patterns and the PCGrimAge epigenetic clock aging. The twins were born between 1880 and 1958 and data on reproduction was collected in 1975 when the twins were between 17 and 95 years old and followed through 2020. The period of the blood collection was 1994-2020 and the age range of the participants at blood drawing was 36-87 years.

The paper has a very bold title: "Human reproduction comes at the expense of faster aging and a shorter life", indicated that the associations demonstrated are causal.

Furthermore, the study is mainly a confirmation of the well-known U-shaped association between parity and mortality risk. The title could therefore be expanded to include the sentence: "...and the absence of human reproduction also comes at the

expense of faster aging and a shorter life". The study also demonstrates that earlier timing of childbirth is associated with a negative aging profile. In other words, it is the large middle group (both in terms of number and timing) that have the most favorable profile in terms of survival and epigenetic clock, which is to be expected.

Although the data and the analyses are of high quality, a major concern is that the twins included in the study are born between 1880 and 1958 – a nearly 80-year-long period. This makes the interpretation of the study very difficult, also in terms of the confounding control. For example, it is probably a very different covariate to have seven years of lifetime education for a woman giving birth in the year 1900 compared to a women giving birth in the year 1980. It is not clear that the historical birth cohort variable is able to accounts for such challenges.

Furthermore, it is remarkable that more than a third of the total sample of women do not have any children. This can be explained by the fact that the study population is between 17 and 95 years old, so the "no children group" consists of a combination of young women/girls who have not started trying or succeeded in getting children and a group of mainly middle-aged and elderly who have not had any children. Whether it is meaningful to combine these groups into one is highly questionable.

Furthermore, the latent class model including six classes defines both the very big group of women with no children and the very small group of women with a high average number of children (6.7). The remaining groups, which consist of the majority of participants, all are within a band from 2.0 to 2.4 children with the timing being the main difference.

So, basically, all the analyses hinge on the two extreme groups that form the basis for the very bold claims in the title. The middle groups (maybe with the exception of class 1 of very early births) show very little differences in survival and aging profile when confounders are adjusted for. As the no children group is a mixed group, and the class 6 is a small group with a very high number of children, the interpretation of the association between human reproduction and aging/survival is very difficult regarding the "typical" range of human reproduction. On top of this, there are difficulties in the interpretation of confounder control over the nearly eight decades of birth cohorts included.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

The revision has addressed the concerns raised by the referees to the extent that the structure of the data set allows, and the limitations of the data set and analysis are now more clearly communicated.

Reviewer #2

(Remarks to the Author)

The authors have mostly addressed my comments. In one case, i am concerned about the statistical approach. In their response to review, they present their random effects model as accounting for factors shared by siblings within a family. This is incorrect. In fact, any family-level factor that contributes to the outcome that is not already explained by the included regressors will BIAS the effect estimates of interest. We should be particularly concerned about such bias in the case of this paper because many factors shared by siblings may influence both fertility and biological aging. Instead of using a random effects model, the authors should be using what economists refer to as a fixed effects model (sometimes called a within-group regression; see here <https://www.sciencedirect.com/topics/social-sciences/fixed-effects-model>) that focuses the hypothesis test on differences between siblings in a family. There are a few ways to implement this. But the easiest way to think about it is as a "Sibling Difference" design where sibling differences in biological age are regressed on sibling differences in fertility.

Reviewer #3

(Remarks to the Author)

The authors have voluminously responded to previous reviewer comments. It is strongly suggested to just put these limitations into the manuscript rather than inching up to them, writing 4 paragraphs in response, and then disregarding in the text.

Some points for perspective:

Pregnancy is a well-known physiological stressor. As noted by multiple reviewers, this isn't a new finding or surprising. Well documented phenomena across populations.

The sub-sample of epigenetic measures taken between 1994-2020, along with the table 1 data on the underlying population could be a case-study for selection bias, and plausible avenues for why what is observed in the data is observed. Based upon the voluminous responses to the reviewers, consulting a scientist who understands this concept would have been a better use of everyone's time. In short, there appears to be survival based requirement to be included, along with a response requirement. The reader doesn't know the baseline denominator of potential subjects for the baseline assessment. The epigenetic data was a convenience sample and not representative of the overall cohort.

While the goal of the research may indeed have a causal lens, the methods applied are not estimating a causal effect, but the language in the text throughout implies causality. This was not changed, despite reviewer comments.

The Cox regression survival analysis does not define how follow up was calculated or what time 0 was, making it impossible to interpret. It takes a puzzle master to figure out when the various covariates were measured/included as well, and still unclear how they align.

BMI is not a lifestyle factor.

Yes, the application of latent class modeling to derive "distinct" classes helps statistically, but what any of the classes represent beyond the nulliparous and those with a high number of pregnancies. Despite this, there's nothing really going on except on the tails of the distribution, and despite nulliparous women being at the highest risk compared to whatever latent class 3 is, the text largely ignores this and spends most of the time speculating and interweaving supported theories with evolutionary reproduction factors and focusing the conclusions on this.

Reviewer #4

(Remarks to the Author)

The authors have been very responsive to the reviews and revised the manuscript satisfactorily.

Version 2:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed my comment. I note that the approach the authors opt for is vulnerable to bias in the case that unmeasured factors are correlated with both the outcome and the unit of the random effect (in this case, the family unit). Such factors are almost certainly present -- genetics, unmeasured variability in socioeconomic circumstances, family norms in health behaviors, etc. There is a formal test to determine if random effects are valid in this setting proposed by Houseman (e.g. see here <https://www.sciencedirect.com/topics/social-sciences/hausman-test>). It almost always rejects the random effects approach, though. An alternative might be to evaluate the consistency of results using alternative methods to account for the non-independence of observations within families: clustering standard errors at the family level, generalized estimating equation (GEE) models. At minimum, the authors should acknowledge the vulnerability of their analysis to this type of confounding.

Reviewer #3

(Remarks to the Author)

The authors have largely responded to everything and updated the manuscript. One point remains which needs updating, R 3.2:

The following statement should be removed from the text: "Nevertheless, studies utilizing epigenetic clocks within this dataset have produced results that align well with other population studies⁴⁴⁻⁴⁶, suggesting that major selection effects are unlikely."

This is incorrect and not an approach or valid way to address or assuage the selection bias.

Version 3:

Reviewer comments:

Reviewer #2

(Remarks to the Author)

The authors have addressed my comments

Reviewer #3

(Remarks to the Author)

The authors have sufficiently addressed the review comments this reviewer raised.

Open Access This Peer Review File is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

In cases where reviewers are anonymous, credit should be given to 'Anonymous Referee' and the source.

The images or other third party material in this Peer Review File are included in the article's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

REVIEWER COMMENTS

Reviewer #1 (Remarks to the Author):

This paper provides a new type of evidence from a large data set of monozygotic and dizygotic twins that supports the evolutionary theory of aging in some detail. I am comfortable with the statistical analyses, but I would like to suggest a few changes to the citations:

R1.1: This paper also provides genetic evidence, of a different kind, supporting the evolutionary theory of aging:

Long, Erping; Zhang, Jianzhi. Evidence for the Role of Selection for Reproductively Advantageous Alleles in Human Aging. *Obstetrical & Gynecological Survey* 79(3):p 165-166, March 2024. | DOI: 10.1097/01.ogx.0001010448.83325.67

We thank the reviewer for pointing out this study which indeed supports our hypothesis in a very relevant way. We now refer to it in our Main Text, for example, at line 100:

Finally, it has so far been challenging to assess the influence of confounding factors, such as early life, socio-economic background, or genetic determinants, that could affect both reproduction and lifespan: for example, Long and Zhang recently identified 98 variants that promoted reproduction but decreased survival¹².

R1.2: Line 47: This is a better reference to the current state of evolutionary medicine (your reference 10):

Stearns, S.C. & Medzhitov, R. 2024. *Evolutionary Medicine* 2nd Edition. Oxford University Press. 400 p. ISBN-10: 0192871985

We have now updated the reference to this newer edition.

R1.3: Line 54: This paper provides additional evidence of a U-shaped mortality curve:

Wang, X., Byars, S.G., & Stearns, S.C. 2013. Genetic links between post-reproductive lifespan and family size in Framingham. *Evolution, Medicine, and Public Health*. 2013. doi: 10.1093/emph/eot013

We thank the reviewer for this suggestion, and have now incorporated this citation into our introduction and discussion as reference 11.

R1.4: For future research: are at least some of the loci identified by Long and Zhang also identified in the epigenetic data sets?

We thank the reviewer for this insightful idea. We sought to investigate whether the methylation status near the 98 SNPs that were found to have antagonistic pleiotropy effects in Long and Zhang (2023) would associate with the number of children. We tested this in a dataset of 2,636 people, and methylation values from all CpG-sites within 1000 basepairs up- or downstream of the SNPs (in total 281 CpGs near 93 SNPs in our methylation data). Interestingly, we did find associations between the methylation and number of children, especially in chromosomes 1, 2, and 12, with genes associated to i.a. age at menarche and menopause.

We think these findings are very interesting and possibly worth further investigations, of which we would be happy to discuss with the reviewer after the review process.

Reviewer #2 (Remarks to the Author):

This article reports analysis of associations of reproductive history with survival and biological aging in a large Finnish Twin cohort. The main findings are that women who have had more pregnancies tend to be biologically older, as measured by the GrimAge epigenetic clock, and have shorter lifespans as compared with women who have had fewer pregnancies. In addition, they find that early childbearing or being nulliparous (never having children) are also associated with these phenotypes of accelerated aging.

This paper makes a contribution to the literature on reproductive history and biological aging. George Williams theorized that aging represented a tradeoff between reproduction and somatic maintenance. This theory was later refined by Thomas Kirkwood. Here, the researchers report evidence expanding on findings by Ryan et al. There is longstanding epidemiological evidence of shorter lifespans and higher morbidity among women with more children or more pregnancies. More recently, Ryan et al. along with others have reported evidence of accelerated biological aging as measured by epigenetic clocks. This paper adds to that literature by including an analysis of sibling differences that partly controls for genetic confounding of associations between reproductive history and biological aging.

I have a few comments that, being addressed, would enhance the impact of the article.

R2.1: The authors quantify reproductive history using a latent class model to assign women to categories based on their probability of live birth in each of a series of 3-year time windows. The intuition for why this approach is superior to simply counting the number of live births or alternative specifications of fertility history is not clear from the text. The authors should offer more explanation for why their chosen method is superior to more straightforward approaches. They should also report results from simple tabular approaches somewhere in the article.

We thank the reviewer for this important suggestion. In fact, this suggestion accurately reflects the genesis of this study: we started out with a tabular approach to the question, and later turned to LCA due to the limitations of the tabular approach. We have now restored these more direct analyses in the manuscript, and comprehensively discuss the rationale and relevance of the LCA approach in this context.

The results of the tabular approach are presented as Extended Data Figure 7, and as Supplementary Information (Methods and Results).

Main text (manuscript):

Line 61:

Interestingly, earlier reproduction has also been linked to a shorter lifespan, which may reflect the reproduction-survival trade-off irrespectively of the number of offspring¹². Yet, to our knowledge, most studies have only focused on effects of age at first reproduction on lifespan^{5,10,11,12}, and have not assessed the timing of each consecutive childbirth. But a reproductive history (including both the number and the timing of childbearing events) is highly multivariate: while classical modelling approaches are based on selected indicators (typically offspring count, age at first reproduction or average interbirth interval), they need to balance parameter space, collinearity between parameters, and accuracy in the representation of reproductive trajectories. Mixture modeling methods, such as latent class analysis (LCA)¹³ classify reproductive histories into categories based on timing and frequency, capturing nuanced patterns while reducing overfitting and minimizing oversimplifications inherent in tabular methods.

Line 112: *Based on a dataset consisting of 14,931 Finnish twins⁷ allowing us to account for shared genetic and environmental effects, we assess the effect not only of the number of offspring, but also the timing of each childbirth. To that aim, we use a dimension-reduction approach based on latent class analysis (LCA): briefly, we identify K “typical” reproductive trajectories based on a very large reference sample (N=14,931) in our focal population (19-20th century Finland), and describe each individual as a weighted mixture of these K components. This allows for (1) the representation of any individual trajectory as a fixed-length vector of weights of size K, and (2) for analysis of the outcomes of each of the K typical strategies without attempting to separate the interlinked effects of e.g. age at first reproduction and total number of offsprings, and (3) to explore whether our study participants can be differentiated into subgroups based on their*

pregnancy history, simultaneously providing valuable insights into the reproductive behavior of Finnish women born since the 1880's until the mid-20th century.

Line 212: *These results are also consistent with tabular approaches, as we modeled survival and epigenetic age acceleration as a function of parity and age at first child, where we found increased mortality risk and accelerated aging for the nulliparous women, and women with five or more children (See Supplementary methods and results, and Extended Data Figure 7).*

Supplementary methods:

3.4 The effects of parity and age at first child on later health

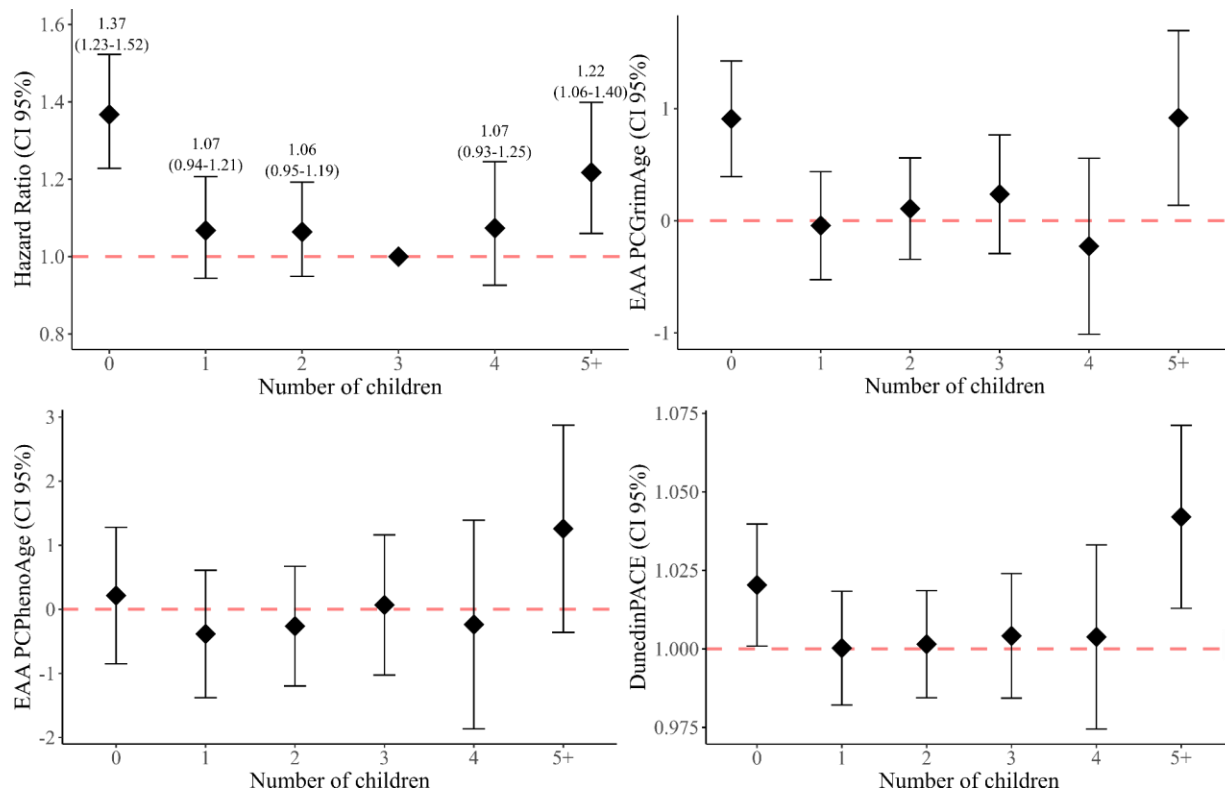
The effects of the number of children and age at first child were explored using categorical child count with five levels (one, two, three, four and five or more children), and age at first child, respectively. All the models were run without imputation of missing covariates. First, we modeled survival with mixed Cox proportional hazards models with adjustments for left truncation, alcohol use, smoking, bmi and education, using the same modeling methods in R as for the latent class analysis. Women with three children were used as a reference group for the survival analysis. Second, we modelled epigenetic age acceleration with linear mixed effects models in R using the package lme4⁷⁶. These models were adjusted with chronological age, alcohol and tobacco use, and BMI at the time of blood sampling and lifetime education. All of the above models included Family ID as a random effect to account for related individuals. The estimates are reported using estimated marginal means with the package emmeans⁷⁷.

Supplementary results:

Parity, age at first child, and later health

Our latent class results are consistent with tabular approaches, as we modeled survival and epigenetic age acceleration as a function of parity and age at first child using models with same covariates as in the fully adjusted models (see Supplementary Methods). We observed increased mortality risk for the nulliparous women (HR=1.37, CI 95% 1.23-1.52), and women with five or more children (HR=1.22, CI 95% 1.06-1.40) compared to the women with three children (used as a reference), yet no other differences between parity groups were observed (see Extended Data Figure 7). Also, the women with five or more children and the nulliparous women had accelerated aging estimated by PCGrimAge and DunedinPACE (see Extended Data Figure 7). Later age at first child was associated with slightly decreased mortality risk (HR=0.98, CI 95% 0.97-0.99) and lower DunedinPACE of aging (Estimate=-0.0015, p=0.0247) in our unadjusted models, yet in lifestyle adjusted models this effect was statistically nonsignificant.

Extended Data:



Extended Data Figure 7. Models for the effect of the number of children on survival and epigenetic age acceleration. A. Mortality hazard ratios with 95% confidence intervals for the number of children from a Cox proportional hazards model for the complete case sample (N=10,975). The model is adjusted with relatedness, left truncation, alcohol use, smoking, BMI, and education. The red dashed line is drawn for reference at the estimate for women with three children. B-D. Estimates for epigenetic age acceleration using PCGrimAge (B), PCPhenoAge (C) and DunedinPACE (D) with 95% confidence intervals for the model without covariate imputation using Full Information Maximum Likelihood (N=855). The epigenetic aging models are adjusted with relatedness, chronological age, alcohol and tobacco use, and BMI at the time of blood sampling, and lifetime education. The red dashed line in the epigenetic models is drawn for reference where the rate of epigenetic aging does not differ from the rate of chronological aging (at 0 for epigenetic age acceleration in B-C, and 1 for DunedinPACE).

R2.2: It's not clear why zygosity is not considered in the adjusted models that include family fixed effects (meaning the models that include dummy variables for each family ID). Perhaps analysis could be stratified by zygosity?

We sought to investigate whether adjusting for zygosity would impact our results. In our original models from the manuscript we have adjusted for family ID as a random intercept, which not only corrects for the shared genetic background of the twins in each pair, but also the common early

life environment. We tested the effect of zygosity by rerunning our survival models with a more elaborate random structure, where family ID is nested within each zygosity group. Table R2.2 below shows the comparison between the original models and the models where zygosity has been incorporated. The estimates for the hazard ratios stay largely the same, and zygosity explains 1-2% of total variance in these models, whereas family ID explains over 28% in the models. Thus, the biological inference from our results is unchanged. These results are in line with Hjelmberg et al. (2019), who found that there are no substantial differences in survival between MZ and DZ twins, and thus we expect zygosity to not play a significant role in our analyses.

Literature:

Hjelmberg, J., Larsen, P., Kaprio, J., McGue, M., Scheike, T., Hougaard, P., & Christensen, K. (2019). Lifespans of Twins: Does Zygosity Matter? *Genes*, 10(2), 166. <https://doi.org/10.3390/genes10020166>

Table R2.2. (not intended for manuscript, just for reviewers). Comparison of adjusted survival models: family-level random intercept vs. family nested within zygosity group. Hazard ratios (HR) with 95% confidence intervals (CI; LL = lower limit, UL = upper limit) from adjusted survival models with and without covariate imputation. The models include two random structures: a family-level random intercept (original model) and a structure where family is nested within zygosity group. Results for the fixed effects (HRs for each class) are presented in the upper section, while variance components for the random effects are summarized in the lower section.

Variable	Imputed adjusted survival models						Non-imputed adjusted survival model					
Fixed part	Original model			Model with zygosity			Original model			Model with zygosity		
Class	HR	LL	UL	HR	LL	UL	HR	LL	UL	HR	LL	UL
0	1.41	1.28	1.56	1.40	1.27	1.55	1.33	1.20	1.48	1.33	1.20	1.47
1	1.10	0.98	1.24	1.10	0.98	1.24	1.07	0.94	1.21	1.07	0.94	1.2
2	1.07	0.96	1.20	1.07	0.96	1.9	1.06	0.95	1.19	1.06	0.95	1.19
3	1 (reference)			1 (reference)			1 (reference)			1 (reference)		
4	1.03	0.91	1.16	1.03	0.91	1.16	1.04	0.92	1.18	1.04	0.92	1.18
5	1.02	0.90	1.16	1.02	0.90	1.16	1.02	0.89	1.16	1.02	0.89	1.19
6	1.16	0.99	1.36	1.16	0.99	1.36	1.16	0.99	1.36	1.16	0.98	1.36
Random part	Original model			Model with zygosity			Original model			Model with zygosity		
Family	0.301			0.298			0.287			0.284		
Zyg/Family				0.016						0.011		

R2.3: It's nice that the authors report the twin correlations for GrimAge, although this is a bit off topic from the paper. If they choose to keep this in the paper, they should report it for the other clocks as well.

The twin correlations for the epigenetic age acceleration and survival analysis are indeed slightly outside the focus of the manuscript, and thus we decided to exclude this part of the manuscript.

R2.4: Some sensitivity analysis testing whether results are consistent for women measured on different DNAm platforms would be useful.

We thank the reviewer for raising this important point. We now recalculated EAA values taking into account the different Illumina array platforms used in this study. Namely, (1) we used only probes that are shared across all platforms to calculate epigenetic age (for PC-clocks, 66,260 CpGs out of the total 78,464 included CpGs, and for DunedinPACE, 144 CpGs out of the total 173 included CpGs), (2) we recalculated epigenetic age acceleration by extracting residuals from a model where the epigenetic age is regressed against chronological age, including an interaction effect (variable slope) between age and platform, as well as adjusting for bisulfite-conversion plate (bs_plate). Here we show the difference in these model formulae (using PCGrimAge as an example):

Previous: $\text{PCGrimAge_EAA} = \text{resid}(\text{lm}(\text{PCGrimAge} \sim \text{age}, \text{data} = .))$

New: $\text{PCGrimAge_EAA} = \text{resid}(\text{lm}(\text{PCGrimAge} \sim \text{age} * \text{platform} + \text{bs_plate}, \text{data} = .))$

This effectively corrects for platform-specific biases that could have been carried through to our EAA estimates. We have now derived all our results from the new model, and made changes to the manuscript accordingly (Figure 1, Extended Data Figure 4). We also tested whether adding a platform term to models that use EAA as a response further accounted for potential residual effects, but this did not improve model fit for PCGrimAge, PCPhenoAge nor DunedinPACE (see Akaike's information criteria in Table R2.4). As adding the platform to these downstream models did not improve model fit, and as we had already adjusted for platform in the EAA calculation step, we report the original models adjusted for the same covariates as previously (relatedness, chronological age, alcohol and tobacco use, and BMI at the time of blood sampling and lifetime education) with these newly calculated EAA estimates used as outcomes (Figure 1 in manuscript).

We have modified the Methods in the text in the following manner to describe these changes:

Main text (manuscript file):

Line 527: *Biological age was determined from methylation beta values using principal component (PC)-based⁹ GrimAge⁸ ("PCGrimAge", reported in the Main Text) and PhenoAge¹⁹ ("PCPhenoAge") as well as DunedinPACE¹⁸ (reported in the Extended Data). We used only the probes that were found in all platforms after preprocessing. Epigenetic age acceleration was determined as the residual of epigenetic age linearly regressed against an interaction between*

chronological age and Illumina BeadChip platform, corrected for bisulfite-conversion plate in our sample of 1,054 participants.

Supplementary Methods:

2.4 Epigenetic aging data

Epigenetic aging

Biological age was determined from DNA methylation beta values using three published algorithms: principal component versions⁹ of GrimAge⁸ and PhenoAge¹⁹, as well as DunedinPACE¹⁸. We used only the probes that were present in all platforms after preprocessing (for PC-clocks 66,260/78,464 CpGs; for DunedinPACE 144/173 CpGs). Epigenetic ages of the above algorithms were calculated in R using the scripts provided in the corresponding publications. Epigenetic age acceleration was determined as the residual of epigenetic age linearly regressed against an interaction between chronological age and platform, and bisulfite-conversion plate in our sample of 1,054 participants.

Table R2.4.A. Akaike's information criteria (AIC) for the covariate-adjusted models with and without additional adjustments for Illumina BeadChip platform. The models without platform ("Original") correspond to the manuscript models with the original covariates.

Outcome	Original	With platform
EAA PCGrimge	18 010.16	19 277.55
EAA PCPhenoAge	19 474.82	20 744.10
EAA DunedinPACE	11 354.19	12 655.76

Reviewer #3 (Remarks to the Author):

Overall: This manuscript reports on an analysis of data from Finland that aimed to examined latent classes of pregnancy history in women accounting for timing and number, and examine these statistically derived classes with mortality risk, and also examine a convenience sample of epigenetic based data in relation to the derived classes. Overall, there are many methodologic considerations that appear unaddressed that limit a clear interpretation of the results for the reader and limit a placing of this research in the context of what is actually a fairly robust evidence base (when considering myriad chronic disease conditions) related to the same exposure. Combined with inappropriate causal language interspersed throughout, and overexuberant interpretation, there is limited enthusiasm in the current form.

Specific Comments

Abstract

R3.2: The language implies that a causal effect was estimated, but the nature of the question and the methods applied to answer it are not estimating an effect, thus “impact” is inappropriate language as are further conclusions beyond the sentence below this statement noting the objective is to examine an association.

We agree with the reviewer that our study does not address a causative effect and thus we have adjusted the text accordingly. (See below R3.4)

R3.3: What does “contemporary” supposed to mean to the reader? What about “early” or “numerous”. This is vague/unclear language for the reader.

We have now removed the word contemporary from the text, and specified the concepts of early and numerous in the abstract. (See below R3.4)

R3.4: The conclusive statement does not reflect the research and its not clear why emotive language is included.

We have now adjusted the text to be less assertive. The entire corrected paragraph can be seen below:

Evolutionary theories suggest a trade-off between resources allocated to reproduction and those allocated to self-maintenance, and predict that higher reproductive output entails a shorter lifespan. This study investigates the link between childbearing, aging and lifespan using data from Finnish twin women. We model the association between reproductive trajectories and survival in 14,836 women, and assess biological aging in a subset of participants (N=1,054) using the PCGrimAge clock, an algorithm trained to predict biological aging and mortality risk from DNA methylation. We identified six distinct reproductive trajectories describing different timing and number of pregnancies. Women with the most pregnancies throughout their lives (mean 6.8, SD 2.4), as well as nulliparous women, had accelerated aging and highest mortality risk. These results support the existence of a trade-off between reproduction, aging and lifespan in modern humans, and provide novel insights into the genetic and lifestyle determinants of healthspan.

Main body-Intro

R3.5: 1st paragraph – causal language not supported by evidence – those are associations noted, not effects. Ironically, the last sentence of that paragraph undercuts/contradicts the earlier notions. Which is it? This is confusing for the reader. As noted, this isn't a new topic/idea and it has strong theory behind it, along with multiple studies supporting the theory.

We have now described the associations more precisely in the following:

*Line 55: Multiple studies have found a U-shaped relationship between reproduction and mortality in humans and other species: both nulliparity and giving birth to many offspring has been **linked to increased** mortality risk^{4,5,11}.*

R3.6: Need some references to support statement that no study has examined age of first pregnancy and age/timing/count of any subsequent pregnancies. This statement also implies that there has been a comparison of approaches in the current study – do the data support this in the current study?

We have now re-analysed our data with the same modeling approach using the number of births and age at first child. The results of this analysis support our latent class analysis results (see Extended Figure 7), and complement the fact that our approach ensures we have not overlooked a signal unrelated to the number of children regarding the timing of childbirth.

We have now added description of the previous findings on the age at first child effect in the following lines:

Line 61:

Interestingly, earlier reproduction has also been linked to a shorter lifespan, which may reflect the reproduction-survival trade-off irrespective of the number of offspring¹². Yet, to our knowledge, most studies have only focused on effects of age at first reproduction on lifespan^{5,10,11,12}, and have not assessed the timing of each consecutive childbirth.

Line 97:

What is more, to our knowledge, previous studies have focused on the number of children or the age at first birth, regardless of the timing of subsequent pregnancies, and thus could not consider the effects of the full childbearing history, including later-life pregnancies^{3,20,21}.

R3.7: What is being communicated re: confounders? This meandering statement is unclear and could be interpreted multiple ways. Have previous studies not accounted for likely confounders?

We have now specified what we mean by genetic confounding in line 100 onwards:

Finally, it has so far been challenging to assess the influence of confounding factors, such as early life, socio-economic background, or genetic determinants, that could affect both reproduction and lifespan: for example, Long and Zhang recently identified 98 variants that promoted reproduction

but decreased survival¹². Ryan et al. controlled for genetic background through a longitudinal study design, where they found that pregnancies between age 25 years to 31 years accelerated epigenetic aging using two epigenetic clocks, yet whether these effects were transient or persisted later in life remains unknown³.

Regarding confounding effects, we kindly ask the reviewer to also see R3.26.

Methods

R3.8: What is the overall N of women in the study?

The overall N was 14,931. In the latent class analysis, we analysed reproductive trajectories using data from all women who had given birth (N=10,783). The survival models were conducted using data from 14,836 participants, who were women from the LCA, and the nulliparous women, who lived past 40 years (N=14,836). The epigenetic analyses were conducted with participants with DNA methylation data (N=1,054).

Line 112: Based on a dataset consisting of 14,931 Finnish twins⁷ allowing us to account for genetic effects, we assess the effect not only of the number of offspring, but also the timing of each childbirth

Line 115: briefly, we identify K “typical” reproductive trajectories based on a very large reference sample (N=14,931) in our focal population

Line 127: Among those women with methylation data (N=1,054), we assess differences in biological age acceleration using the PCGrimAge^{8,9} algorithm.

R3.9: How many were sent this questionnaire in 1967? It appears 17,000+ responded – but its important to know for selection bias purposes, particularly since this wasn’t clearly sampled and was a convenience response. Any conclusion is predicated on understanding the selection of the population under study and what is described would suggests a likely biased sample.

The Finnish Twin Cohort was sampled from the entire Finnish population by sampling pairs of persons born on the same day, in the same local community, of the same-sex and with the same surname at birth in 1974. Twinship was confirmed in the baseline survey in 1975 and by inquiries to local magistrates on whether the twin candidate pair had the same parents. In 1975 and later the response rates were very high, suggesting that the twin individuals are more representative (as individuals) of the whole population than corresponding cohort studies of unrelated individuals with lower response rates. We know from register linkage that the non-responders were more likely to have been hospitalized for psychiatric diseases. In mortality and cancer incidence analyses, the cohort does not differ from the whole population when followed up for 3-4 decades (Kaprio 2013, Skytthe et al, 2019). In multisite molecular genetic analyses that data from the cohort has contributed to over the past 15 years, twin individuals do not differ from other cohorts (of unrelated

individuals). We appreciate the reviewer's diligence in ensuring the integrity of our research, however, we have not identified any clear selection bias in our sampling procedure, and again stress that the Finnish Twin Cohorts have been used in a large number of studies to date, and confronted to data from other cohorts, which should at least partly guarantee the robustness of the study system.

Literature

Kaprio J. The Finnish Twin Cohort Study: An Update. *Twin Research and Human Genetics*. 2013;16(1):157-162. doi:10.1017/thg.2012.142

Skytthe A, Harris JR, Czene K, et al. Cancer Incidence and Mortality in 260,000 Nordic Twins With 30,000 Prospective Cancers. *Twin Research and Human Genetics*. 2019;22(2):99-107. doi:10.1017/thg.2019.10

R3.10: Its unclear how a cohort that includes women born in the dates noted could be considered “contemporary”.

We have now removed the word contemporary and worded the specific birth years in lines 157 onwards: *The ability of our model to robustly and convergently identify these classes suggests that Finnish women born during 1880-1957 can indeed be distinguished by their reproductive histories into distinct subgroups.*

R3.11: It also appears the questionnaire was sent to 9-10 year old twins in 1967. How can a latent class analysis be appropriately classified if using data where its possible a number of participants were not even biologically capable of being pregnant, and unlikely to be so for a number of years?

There may be a slight misunderstanding here: the first questionnaire was conducted in 1975, when the youngest respondents were 18 years old. For the participants born after 1949, information on childbirth was validated from the Digital and Population Data Services Agency Finland. Thus, the youngest women who were included in the latent class analysis, for whom solely questionnaire information was available, were 26 years old, not 9-10 years old. We do agree this prevents analysis of life-long age acceleration and survival patterns in these samples, which are practically right-censored. However, the age-acceleration is measured at the time of the latest reproduction data, and is modelled according to events that predate it - by definition we do not model reproductive choices happening after the blood sampling event, as these have little chance of already being reflected in that blood sample. For survival analysis, we further subsetted the sample to only women who had lived past 40 years (N=14,836 of total N=14,931). We understand that mentioning in the text the year 1967 might have been confusing, as this was meant to solely describe how the cohort was initially founded: The 1975 questionnaire was sent to all twin pairs who had been born in 1880-1957, where both co-twins were alive in 1967. Twins whose cotwin had died 1967-1974 were excluded from the 1975 questionnaire target sample. We have now deleted the mention of 1967 from our Main Methods description in line 473:

The 14,931 study participant women were a subset of the older Finnish Twin Cohort⁷ born between 1880 and 1957, ~~with both co-twins alive in 1967~~ (see Supplementary Methods, Supplementary Results and Extended Data Table 3).

The description of initial screening can still be found in our Supplementary Methods under paragraph “**1. Participants**”.

R3.12: Given the different birth cohorts, it seems results presented by these would be merited to make sound conclusions.

The historical cohorts compiled from the birth years of the participants were used in the analysis to control for possible differences in survival related to the living circumstances in Finland during varying societal and political eras. As we report the results for differences between the reproductive classes after adjusting for these cohorts, our manuscript results are aimed to represent class differences independently of the birth cohort.

We agree that this is an important aspect to consider, and we have elaborated on the inclusion of these cohorts in our manuscript:

Main text (manuscript):

Line 165: *All of our survival models were adjusted for the historical birth cohort to control for differences in living circumstances in Finland during the range of years of birth in our sample (see Supplementary Methods and Results).*

Supplementary information (results):
The effect of historical cohort

The period during which our study participants were born influenced their survival prospects, and this was adjusted for by the historical cohort variable. This effect was independent of age at entering the study, which was adjusted by left truncation in our models (see Supplementary Methods). In our final models that were adjusted for (imputed) lifestyle factors, the participants born during 1899-1917 (Pre-Independence) had a hazard ratio of 1.32 (CI 95% 1.02-1.72) relative to the reference group (b. 1880-1898 Industrial era). The women born during the Post-Independence period (1918-1938), the second World War (1939-1945) and the Cold War (1946-1957) had survival prospects with hazard ratios of 0.96 (CI 95% 0.74-1.25), 0.74 (CI 95% 0.55-1.0) and 0.83 (CI 95% 0.625-0.1.107), respectively, yet these groups did not significantly differ from the reference group.

Table R3.12. (not intended for manuscript, just for reviewers). Hazards ratios (HR) with 95% confidence limits (LL=lower; UL=upper limit) for the different historical cohorts to the reference class (b. 1880-1898 Industrial era) from the lifestyle-adjusted models.

Historical cohort	Imputed model			Non-imputed model		
	HR	LL	UL	HR	LL	UL
1899-1917 Pre-Independence	1.319	1.015	1.715	1.33	1.014	1.742
1918-1938 Post-Independence	0.961	0.741	1.246	0.960	0.734	1.255
1939-1945 World War II	0.739	0.545	1.001	0.726	0.531	0.992
1946-1957 Cold War	0.832	0.625	1.107	0.859	0.637	1.159

R3.13: Its unclear how well classified and valid the exposure is for the latest birth cohort, given they were close to 18 when the data started, as well as hadn't had a chance to have the same exposure/patterns as groups they were compared to or apart of the same latent class.

We kindly ask the reviewer to see the answer above to comment R3.11.

R3.14: A major issue with interpreting the results is that the assessment of many factors is either predicated on survival, or likely meaning something completely different depending on the lifestage (for example, BMI in 1975 – could be 90+ years or 18...)

We agree that the effect of BMI is dependent on the age of the participants at the time of assessment in the survival analyses. To test this, we explored whether BMI would interact with the age of the participants at the start of the follow-up (1975). We compared our original manuscript models (with BMI or squared BMI) with a model including interactions between either BMI or squared BMI on the one hand, and the age at 1975 on the other. These interactions did not improve model fit based on AIC comparison (Table R3.14 below), and thus we conclude that the interaction term

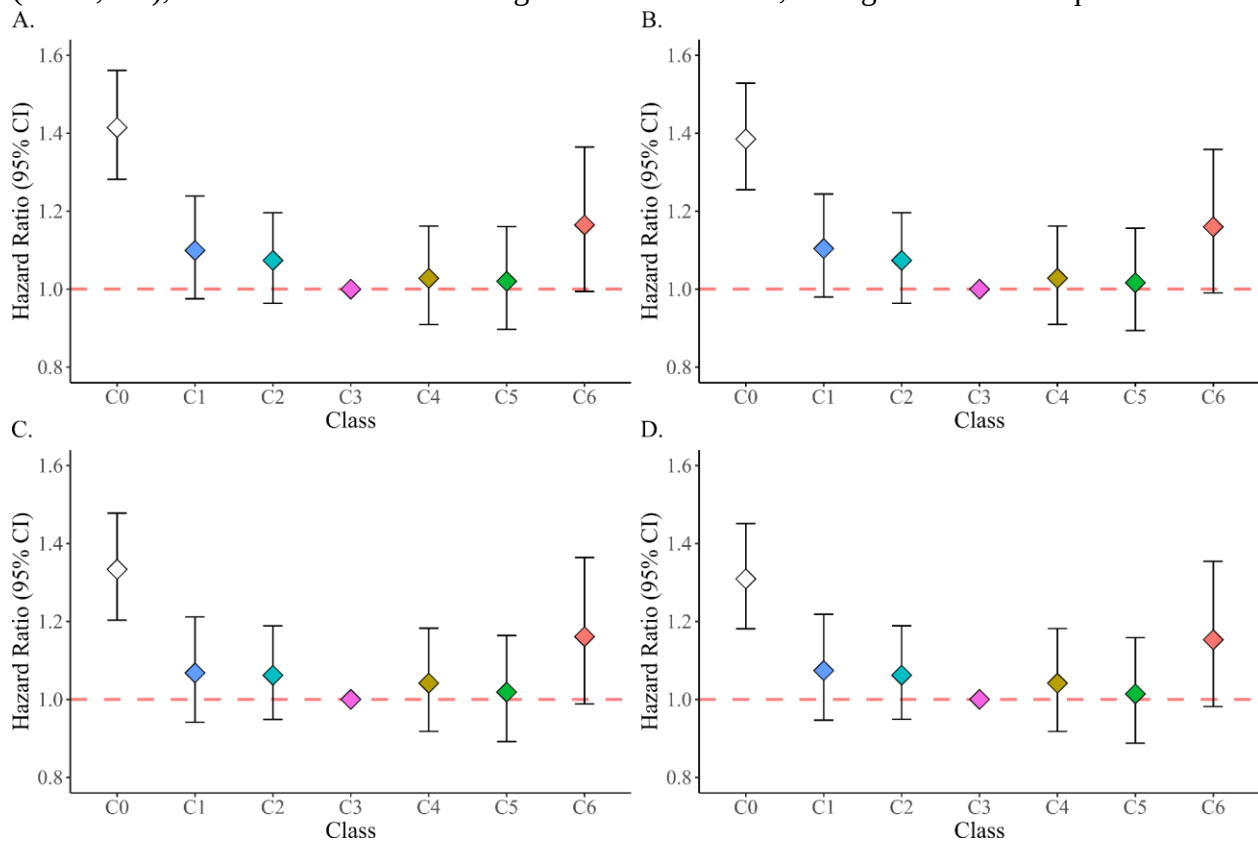
(change in BMI effect with age) does not significantly explain variation in survival or epigenetic aging and influence result interpretation. In Figure R3.14 below, we have also plotted the estimates for each class from the survival models from the original models (A and C) and the ones with the interaction term (B and D). These figures support our conclusion that the interaction with BMI would not change result interpretation.

Table R3.14. (not intended for manuscript, just for reviewers). Akaike’s information criteria (AIC) for the covariate-adjusted models with and without interaction between age and BMI variables (centered BMI and quadratic centered BMI). The models without interaction correspond to the original manuscript models (model with imputed covariates used for survival).

Outcome	Original	Age*BMI
Survival	82 130.56	82 133.9
EAA PCGrimge	18 010.16	19 615.39
EAA PCPhenoAge	19 474.82	21 077.84
EAA DunedinPACE	11 354.19	12 961.83

Plot R3.14. (not intended for manuscript, just for reviewers). Mortality hazard ratios with 95% confidence intervals for each latent class from a Cox proportional hazards model. The dashed red line indicates the reference class used in the survival analysis (class 3). A. Original covariate-adjusted model from the manuscript with imputed covariates (N=14,836). B. Covariate-adjusted model with imputed covariates (N=14,836), and interactions between age and centered BMI, and age and centered quadratic BMI. C. Original covariate-adjusted model from the manuscript with no covariate imputation (N=10,975). D. Covariate-adjusted model with no covariate imputations

(N=10,975), and interactions between age and centered BMI, and age and centered quadratic BMI.



R3.15: When was alcohol assessed?

For the survival analysis, information on alcohol usage was assessed at the start of follow-up in 1975. For the epigenetic age acceleration analysis, we used alcohol usage information that was assessed at the timing of blood sample. We thank the reviewer for pointing out the information on the alcohol assessment time point was not in our manuscript and we have now added it to Supplementary Methods, under subheading **2.2 Contextual data**.

Alcohol

Alcohol use is associated with DNA methylation, epigenetic aging, as well as morbidity and mortality risk^{62,63}. An alcohol use index was compiled from the 1975 questionnaire, first as a continuous variable reflecting average daily use in grams of ethanol⁵⁹, further binned into a categorical variable (“abstainer”, “infrequent” (<1.3g/day), “low” (<25g/day), “medium” (<45g/day), “high” (<65g/day), “higher” (≥65g/day) as described in Zhao et al. (2023)⁶⁴. As heavy drinking was rare among the women in our cohort, the “medium”, “high”, and “higher” categories were combined into a “medium or high” category in the survival analysis. In the epigenetic aging analysis, the alcohol index at the time of blood sampling was binned to three

levels (“non-drinker”=0g/day, “infrequent”<1.3g/day, “frequent”≥1.3g/day) due to the relatively small sample sizes for different categories within each reproductive class.

R3.16: Please report the N of overall and each variable included in multiple imputation. Related, was the data evaluated for the nature of missingness? This is possible and would inform the nature of the imputation, rather than just assuming the most minimal biased approach as reported.

The number and proportion of missing values per each variable are described in our Extended Data Table 3 for both survival data and epigenetic data. We have now included additional details on how missingness was correlated within participants in our Supplementary Methods section 2.3 Handling of missing data (full paragraph pasted below). We also added how we statistically evaluated the nature of missingness.

Supplementary Methods:

2.3. Handling of missing data

Not all participants answered the 1975 questionnaire, and thus had missing values for our four lifestyle covariates (alcohol use, smoking, BMI and education, Extended Data Table 3). For our survival analysis sample (N=14,836), there were 10,975 women in the complete case sample, 164 women with only one missing covariate, 17 women with two missing covariates, 325 with three missing covariates and 3,355 women with all four missing covariates. We evaluated the nature of missingness using the mcar_test from the R package naniar⁶⁹, which showed our four covariates (alcohol use, smoking, BMI, and education) were not missing completely at random (MCAR; Little’s $\chi^2=94.3$, $p=1.13 \times 10^{-9}$). Thus, we assumed the data were missing at random (MAR)⁴². For survival analysis, we used Multiple Imputation (MI) to handle missing covariate values using Mplus software (version 8.2). MI is assumed to produce unbiased parameter estimates and standard errors, as all the available data is utilized to estimate missing variables⁴². We generated 20 datasets (as recommended in Little et al. 2013⁴²) with imputed covariates, and the subsequent modeling was conducted in the same way as the complete case models, by handling multiple imputed datasets with the R package mitools⁷⁰. For comparison, we also ran the survival analyses with the complete case sample without imputation (Extended Data Figure 3). For epigenetic aging data (N=1,054), there were no participants with missing education information. There were 890 women in the complete case sample, 35 women with one missing covariate, 132 women with two missing covariates and 60 women with all three (alcohol use, smoking and BMI) missing. We analysed epigenetic age acceleration with Bolck-Croon-Hagenaars (BCH) approach using Full Information Maximum Likelihood, which takes into account all observed data, including missing data (see section 3.3. Epigenetic aging).

R3.17: How was the epigenetic data sampled- random? Convenience? Some sort of applied sampling approach? When?

We used all the available DNA methylation data that existed for the FTC, resulting in 1,054 women out of our 14,931 total sample. This data has been produced for targeted subgroups and studies on common diseases (including obesity, hypertension, osteoarthritis and macular degeneration) and risk factors (smoking, physical activity to mention a few) based on the questionnaire information from the main surveys of the FTC cohort. Also, we have generated methylation data just to increase sample size, without prior selection based on any phenotype. As there is no dominant basis for selection, we cannot see any major selection effects. In other studies performed on epigenetic clocks in the cohort, we observe results that are very consistent with other population studies. The timing of blood samples varied across participants ranging from age of 36 years to 89 (see Supplementary Methods “**2.4 Epigenetic aging data**”). The overall description of the cohort data can be viewed at our reference number 7:

7. Kaprio, J. *et al.* The Older Finnish Twin Cohort - 45 Years of Follow-up. *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.* **22**, 240–254 (2019).

R3.18: Given the strong confounding effect of factors like age, different generations, and others, plus potential selection biases on who was sampled, how is the reader supposed to interpret the epigenetic data? Shouldn't the reader be interested in the comparison of people of very similar age ranges to make any sort of conclusion?

We agree with the reviewer that an ideal case-control study design to answer this question would involve the recall, at equal age (say 75), of twin pairs born within a narrow time period - such a design would indeed minimise the confounding effect of both age and birth cohort (although not of lifestyle confounders). While we agree this would be ideal, collecting a sufficient sample size structured in this way would involve a logistical and financial effort well beyond the scope of this (and of most) studies - and it would also be subject, this time, to a strong survivorship bias.

The alternative would be to conduct a cohort analysis, with sufficient follow-up time to take into account mortality. When we started the twin cohort in 1975, we had resources to collect questionnaire information on over 31 000 twins, but at that time no resources to take biosamples. In the 1970s and 1980s biosamples were primarily blood serum and plasma samples, while cells/DNA was not collected. The technology for large-scale genotyping evolved in the 1990s such that we decided to start collecting DNA, as a collaboration between our group (professors Koskenvuo and Kaprio) and the genetics lab at the Institute for Public Health, Finland headed by professor Leena Peltonen, already then a well-known molecular geneticist of Mendelian diseases.

Her lab moved into common diseases of complex etiology, using linkage analyses. This period lasted into the new century, when genome-wide arrays became available and our first GWAS studies were conducted. The DNA collected from the mid-1990s onwards from the twin volunteers has been the basis for the current analyses of epigenetic age. So, the ideal cohort analysis of having DNA sampled at baseline (in 1975) from a narrow range of birth years and then following the women through their reproductive years and later for mortality is not logistically possible. The current analysis is in our opinion a very good approximation of the ideal cohort design using the resources available to us.

In this study, we do our best to account for birth cohort and age at sampling in both survival and EAA models. In the epigenetic age acceleration analyses, the participants have been born at a narrower time span (between 1909 and 1957). We controlled for the age effect by 1) focusing on epigenetic age acceleration, which is the residual of the epigenetic age from the actual age at blood sampling, and 2) including the chronological age as a covariate in the analysis. Regarding possible sampling bias, see R3.17.

We have now added the range of birth years as a row in our Extended Data Table 3, to aid the reader to make inference of the possible factors influencing our results.

We have now added text to acknowledge this limitation in manuscript line 312 onwards:

Secondly, the wide range of birth years in our sample (1880-1957 for survival, 1909-1957 for EAA) may influence the estimation of survival and aging patterns. We addressed this by controlling for cohort entry differences through left truncation in survival analyses and by including historical birth cohort as a covariate. The narrower birth year range in the epigenetic age acceleration (EAA) analyses (1909–1957) reduces this concern, with further adjustments using residual-based age acceleration measures and chronological age as a covariate. Furthermore, our study uses a twin design, which mitigates the challenges posed by the wide birth year range by controlling for genetic factors and shared early-life environments. This approach reduces biases from cohort differences and variations in early-life conditions, which are known to significantly influence later-life health and well-being⁴¹. Nonetheless, future research could benefit from more focused cohort designs with narrower age ranges to refine these analyses and minimize potential biases associated with generational differences. However, including participants across a broad temporal span has also allowed us to assemble a large study sample, enhancing statistical power and enabling robust examination of reproduction, aging, and lifespan.

R3.19: It appears the survival data was predicated on participating in the 1975 assessment – what is the N relative to the beginning sample?

Yes, the start of follow-up was the year 1975 for the survival analysis. The total sample size of our study was 14,931 study participants. The survival models were conducted using data from 14,836

participants, who were women from the LCA, and the nulliparous women, who lived past 40 years. Of these 14,836 survival analysis participants 5,225 died during our study window (Extended Data Table 3: *Age at death*).

Survival analysis

R3.20: What is the sample size and selection for this analysis? The truncation summary suggests that there was a need to participate in 1967 and be alive in 1975, which suggests that this was the start of follow up or time 0?

The sample size for the survival analysis was 14,836. We explain the sample selection in the Supplementary information 3.2 - Survival: *“Study participants only entered the study provided they were alive in 1967, we therefore used 1967 as an entry date for left truncation in our model. Right-censored date of exit was either death, emigration or end of follow up, whichever came first.”*

R3.21: The follow up time calculation suggests that no participant was lost to follow up – is this the case?

As we have described in our Supplementary methods section **2.5 Mortality**: *The survival time for each woman was defined in years until the date of emigration, death, or end of follow-up, whichever came first, as well as section 3.2 Survival: Right-censored date of exit was either death, emigration or end of follow up, whichever came first.*

The Finnish population registration system is very comprehensive as legally all moves from one address to the next are reported to the population register, which then updates other authorities such as the tax offices, etc. Moves abroad are registered, and follow-up ends at emigration, death or end of follow-up. Other losses to follow-up such as disappearance without any recovery of a body are very rare. Follow-up for all practical purposes is complete, especially when dealing with volunteer participants in health surveys, who have family (i.e. their cotwin in our case).

R3.22: For the modeling, for the youngest participants, the variable of lifetime education at the 1975 date is likely to misclassify this confounder.

As described in Supplementary methods: 2.2. *Contextual data: Education*, the lifetime education -variable was compiled from answers about attained educational level in questionnaires from 1975 and 1981. Our study participants were thus at least 24 at the time of reporting the education. For the purpose of this study, anyone still engaged in their education at that age would be automatically assigned to “tertiary”-education category. Our three-class education category refers to primary=7 years, secondary=10 years, and tertiary>10 years. Thus, all participants who have completed more than upper elementary school were assigned the category “tertiary”. In Finland, it is very typical that higher education is completed relatively shortly after primary education, and thus likely all

our participants had already continued their education if going to high school, before reproduction. For persons born in the 1950s and earlier, compulsory education was six years, with selection into more academic paths after four years of primary school. The school reforms in the early 1970s led to compulsory and uniform education for 9 years (from age 7 to 16). Only recently was schooling to age 18 made compulsory. We do agree that we might mis-classify a handful of very atypical cases (tertiary education initiated at an age older than 24, but before having children), but these are unlikely to have any statistical impact on our results.

R3.23: Why was C3 the reference group? There is no logic or rationale provided.

We selected class 3 as the reference group for the survival analysis because it corresponded to the average age at childbirth in our total sample: 27.3 (SD 5.7), with 24.4 (SD 4.9) and 29.8 (SD 5.7) for first and last childbirth (see Extended results: Descriptive results), while class 3 had the most probable timing of reproduction at ages 27-29 years. We now clarified this in the manuscript. Class 3 also had the lowest hazard ratio in the survival analysis, which validated our decision to use this class as the reference group, as it would make result inference easier for the reader. However, we also underline that this reference is relative - any class could be used as a reference without changing our conclusions.

We have now added the description for the reference group in our Supplementary Methods section **3.2 Survival:**

We used class 3 as the reference class in our survival analyses to reflect a typical class for reproductive trajectory, as this class had the highest probability of childbirth at age 27-29, while the average age at childbirth was 27.3 in our total sample (see Extended Results).

Results, etc..

R3.24: Supplemental doesn't really suggest any higher risks, or higher epigenetic based age calculations except perhaps in c6 class, but also in nulliparous group. The main results, with imputation, are largely suggestive except for a signal in the nulliparous group. How does this support the over the top conclusions noted by the authors? This is misleading.

For PCGrimAge (Figure 1) and DunedinPACE (Extended Data Figure 4), we found accelerated aging for class 6, the highest reproductive output group (Extended Data Table 4). In line with this observation, in our additional analyses where we model survival and epigenetic age acceleration by the number of children, we find accelerated epigenetic aging for the women with 5 or more children (Extended Data Figure 7). We find this to support our conclusion regarding the life-history-theory-based hypothesis on the reproduction-aging trade-off. For PCGrimAge, we found accelerated aging for the nulliparous group, yet the same result was not replicated for DunedinPACE in our lifestyle-adjusted models. Regarding the timing of reproduction, in our new

analyses suggested by Reviewer 2, for PCGrimAge, PCPhenoAge and DunedinPACE, we find accelerated aging for the early mothers (C1), yet this effect does not remain significant after lifestyle adjustment for PCGrimAge and PCPhenoAge. Discrepancies between clocks are expected due to the different calibration strategies and training populations used when building these clocks: PCGrimAge is trained to predict time-to-death (the reason why we use this in our main analysis), while DunedinPACE is trained to measure biological deterioration over time, and PCPhenoAge is trained to reflect biological age by incorporating clinical biomarkers. Based on these results, we have now toned down the conclusions regarding the age at childbirth in our abstract, summary and main text.

R3.25: What is the reader supposed to make of the differential distribution of participant exposure classes by self-report, v. selected epigenetic based measures?

We agree it is possible that some participants may not have answered accurately to the questionnaires regarding the exposures, which is a common problem in cohort-based studies. **For the number and timing of pregnancies**, we have validated this information by retrieving the exact birth dates of each child from the Digital and Population Data Services Agency in Finland for all participants born between 1950-1957. **The BMI** information from the questionnaires has been previously shown to correlate well ($r(\text{Pearson})=0.95$, $\text{CI}=0.94-0.96$) with measured values in the same cohort (our reference 40, Tuomela et al. 2019). **For smoking**, we show below in Figure R3.25 the correspondence between 1) self-reports as the questionnaire smoking status and smoking *pack-years*, which is a measure of cumulative smoking exposure, calculated as (packs smoked per day) $\times$ (years as a smoker) 2) DNA methylation at the AHRR locus that has been shown to be the strongest singular CpG to predict smoking (Grieshaber et al. 2020), and 3) clinical measures of blood cotinine. These plots show that the self-reported smoking variables correspond to methylation data even at the resolution of one CpG, and clinical measures of blood cotinine. **For alcohol**, unfortunately we could not do similar validation for the questionnaire data, as we have no clinical measures reflecting actual alcohol usage. However the overall reliability of self-reports for BMI and tobacco use suggest alcohol usage may have also been reported accurately. We also note that in our analyses, alcohol use is binned in three broad categories (non-drinker, infrequent or frequent drinker) to avoid excessive reliance on these self-reports. Further evidence for the validity of the alcohol assessment comes from the very strong association of alcohol consumption categories with deaths from alcohol-related disease, see supplemental table e4 in Hublin and Kaprio (2023). We acknowledge the possible unreliability of the questionnaire data is still a limitation of our study. We have now added a mention of the smoking data validation to Supplementary Methods: **2.2 Contextual data: Smoking**, and a description of this limitation to our Main Text.

Supplementary Methods: 2.2 Contextual data: Smoking

We validated the reliability of self-reported smoking information by calculating correlations between self-reported smoking pack-years (calculated as packs smoked per day multiplied by years as a smoker), preprocessed AHRR DNA methylation at locus cg05575921, which has been previously found to be hypomethylated in heavy-smokers⁶⁰, as well as blood-based cotinine measures⁶¹. The Pearson correlation coefficient between self-reported pack-years and preprocessed cg05575921 methylation was -0.48 ($p < 2.2 \times 10^{-16}$), whereas the correlation between pack-years and blood cotinine was 0.61 ($p < 2.2 \times 10^{-16}$).

Main text (manuscript):

Line 307: **Limitations**

Our study has some limitations. Firstly, we rely on self-reported smoking status, BMI, alcohol usage and information on the number and timing of childbirths. However, we have validated the information on childbirth for women born since the year 1950 (over 40% of women in our study sample) from the the Digital and Population Data Services Agency of Finland, and self-reported anthropometric measures have been previously shown to correspond accurately to clinically measured information⁴⁰ (see Supplementary Methods),

Literature:

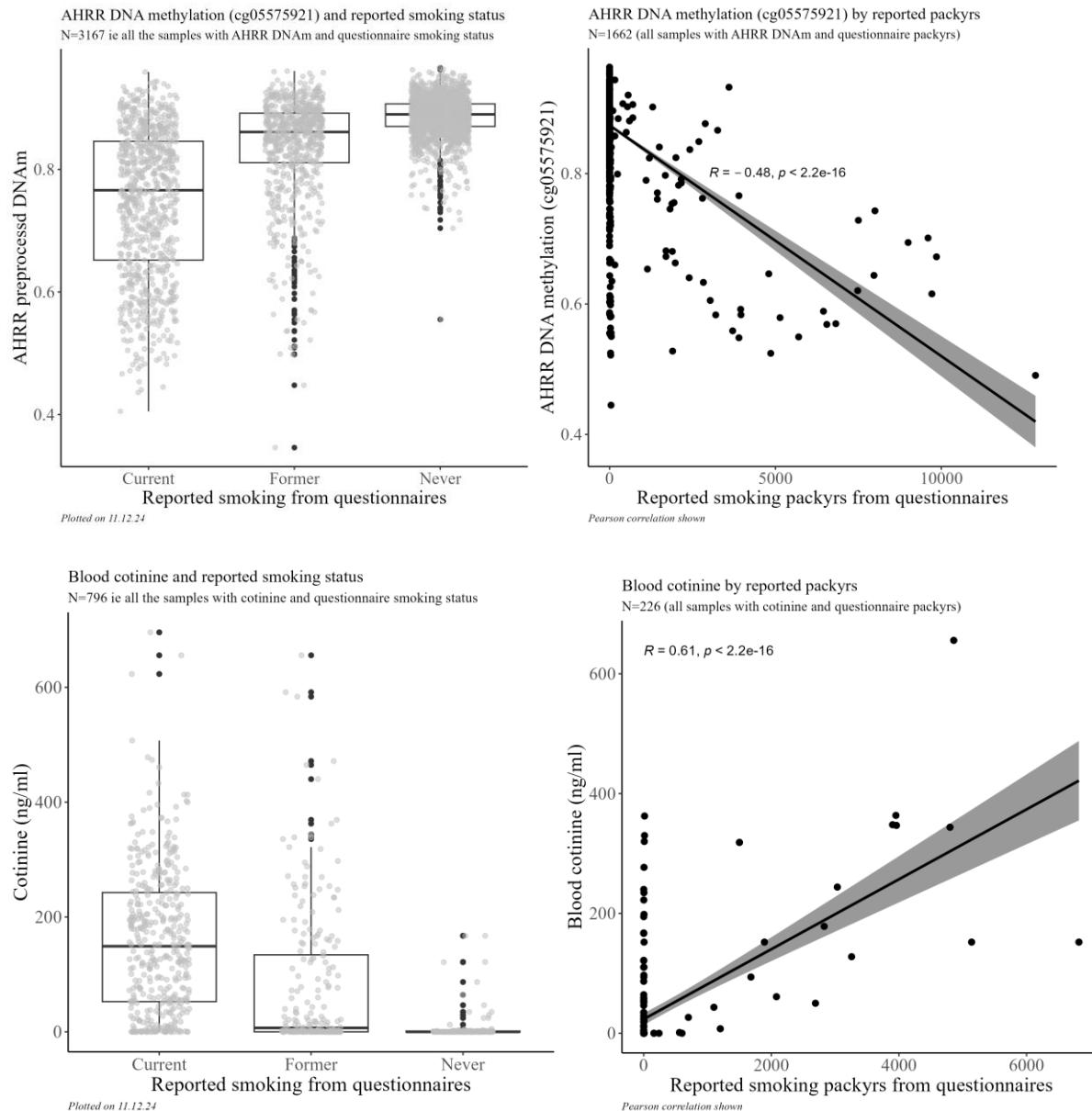
Tuomela, J. *et al.* Accuracy of self-reported anthropometric measures — Findings from the Finnish Twin Study. *Obes. Res. Clin. Pract.* **13**, 522–528 (2019).

Grieshober, L., Graw, S., Barnett, M.J. *et al.* AHRR methylation in heavy smokers: associations with smoking, lung cancer risk, and lung cancer mortality. *BMC Cancer* 20, 905 (2020).

Bollepalli S, Korhonen T, Kaprio J, Anders S, Ollikainen M. EpiSmokEr: a robust classifier to determine smoking status from DNA methylation data. *Epigenomics*. 2019 Oct;11(13):1469-1486. doi: 10.2217/epi-2019-0206. Epub 2019 Aug 30. PMID: 31466478.

Hublin C, Kaprio J. Chronotype and mortality - a 37-year follow-up study in Finnish adults. *Chronobiol Int*. 2023 Jul 3;40(7):841-849. doi: 10.1080/07420528.2023.2215342. Epub 2023 Jun 15. PMID: 37322846.

Figure R3.25. (Only for reviewers): Correspondence between AHRR gene methylation, blood cotinine, and reported smoking status and *packyrs*.



R3.26: Residual confounding due to lack of more comprehensive measurements is a likely strong potential explanation for the findings.

We agree that some unmeasured lifestyle related factors might still confound the results, as our covariates are based on self-reported questionnaire data. However, our study differs from previous research by utilising a twin design, which allows us to control for genetic confounders as well as factors related to early life circumstances, as the twins share their upbringing environment. Previous research has shown that early-life environments significantly influence later-life health and lifestyle outcomes (Carr 2019). Childhood experiences, including socioeconomic conditions

and family stability, can affect long-term health through mechanisms such as stress regulation and behavioral patterns (Carr 2019). These early factors contribute to risks for chronic diseases and impact mental health, social relationships, and overall well-being across the lifespan (Carr 2019). Furthermore, we adjust our models with smoking, alcohol usage, age, education and BMI, which are likely among the strongest confounders regarding studies on aging and health (see Supplementary Methods for covariate-specific rationale). We have also found these variables to associate with later health in the same cohort of this study (Kaprio et al 1987, 2019; Latvala et al. 2014; Heikkinen et al. 2022, Stephenson et al. 2021; Lundgren 2022). Previous studies on parity and later health have most often adjusted for socioeconomic indicators and in some cases smoking (Grundy & Tomassini 2005; Wang et al. 2013; Ryan et al. 2018, 2024; Barclay & Kolk 2019), while we have found only one study controlling for all the same covariates as we in the current study (BMI, smoking, alcohol use and education; Kresovich et al. 2019). As previous studies have deemed these adjustments sufficient, and as the differences between our unadjusted and adjusted models indicate that our covariates explain variation in later health, we propose these adjustments to largely account for confounding lifestyle factors.

Literature:

Carr D. Early-Life Influences on Later Life Well-Being: Innovations and Explorations. *J Gerontol B Psychol Sci Soc Sci.* 2019 Jun 14;74(5):829-831. doi: 10.1093/geronb/gbz054. PMID: 31111921.

Latvala A, Rose RJ, Pulkkinen L, Dick DM, Korhonen T, Kaprio J. Drinking, smoking, and educational achievement: cross-lagged associations from adolescence to adulthood. *Drug Alcohol Depend.* 2014 Apr 1;137:106-13. doi: 10.1016/j.drugalcdep.2014.01.016. Epub 2014 Jan 31. PMID: 24548801; PMCID: PMC3964260.

Kaprio, J. et al. Genetic Influences on Use and Abuse of Alcohol: A Study of 5638 Adult Finnish Twin Brothers. *Alcohol. Clin. Exp. Res.* 11, 349–356 (1987).

Kaprio, J. et al. The Older Finnish Twin Cohort - 45 Years of Follow-up. *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.* 22, 240–254 (2019).

Heikkinen, A., Bollepalli, S. & Ollikainen, M. The potential of DNA methylation as a biomarker for obesity and smoking. *J. Intern. Med.* 292, 390–408 (2022).

Stephenson, M. et al. Associations of Alcohol Consumption With Epigenome-Wide DNA Methylation and Epigenetic Age Acceleration: Individual-Level and Co-twin Comparison Analyses. *Alcohol. Clin. Exp. Res.* 45, 318–328 (2021).

Lundgren, S. et al. BMI is positively associated with accelerated epigenetic aging in twin pairs discordant for body mass index. *J. Intern. Med.* 292, 627–640 (2022).

Grundy, E. & Tomassini, C. Fertility history and health in later life: a record linkage study in

England and Wales. *Soc. Sci. Med.* **61**, 217–228 (2005).

Wang, X., Byars, S.G., & Stearns, S.C. 2013. Genetic links between post-reproductive lifespan and family size in Framingham. *Evol. Med. Public Health* **1**, 241-253 (2013).

Ryan, C. P. et al. Reproduction predicts shorter telomeres and epigenetic age acceleration among young adult women. *Sci. Rep.* **8**, 11100 (2018).

Ryan, C. P. *et al.* Pregnancy is linked to faster epigenetic aging in young women. *Proc. Natl. Acad. Sci.* **121**, e2317290121 (2024).

Barclay, K. & Kolk, M. Parity and Mortality: An Examination of Different Explanatory Mechanisms Using Data on Biological and Adoptive Parents. *Eur. J. Popul.* **35**, 63–85 (2019).

R3.27: Is the interpretation of the survival data as “longevity” appropriate? These were never compared w/in the birth cohorts, and probably can’t reliably be compared except in the latest (cold war) cohort.

We kindly ask the reviewer to see our response to comment R3.18 regarding the assessment of different birth cohorts.

We have now replaced the word longevity in our manuscript lines:

- **211:** *Links between reproduction and **later health***
- **256:** *Simultaneously, higher education is known to be associated with healthier lifestyle and increased survival while delaying pregnancy, likely explaining the decreased **mortality risk** and slowest aging in the classes of later motherhood*

R3.28: Given the unclear comparability considerations around the epigenetic based data, it’s not clear what they add.

We are not sure what the Reviewer is referring to here. We assume this comment points out the discrepancy between different epigenetic clocks. We used three epigenetic clocks in this study that all measure different aspects of biological aging. Discrepancies between clocks are expected due to the different calibration strategies and training populations used when building these clocks:

- PCGrimAge: This clock is calibrated to predict time-to-death, incorporating DNA methylation based estimates of eg. plasma proteins that have previously been associated with mortality or morbidity. It is particularly suited for analyzing mortality risk, which is why it serves as the primary measure in our analysis. (Lu et al. 2019)
- DunedinPACE: Unlike PCGrimAge, DunedinPACE focuses on the rate of biological deterioration over time. It captures dynamic aging processes and provides a measure of the pace at which aging occurs, often aligning closely with PCGrimAge due to their shared emphasis on biological decline and mortality risk. (Belsky et al 2022)

- PCPhenoAge: This clock estimates biological age by integrating DNA methylation data with clinical biomarkers such as glucose and inflammatory markers, and has been focused to predict overall healthspan (eg. also Alzheimer's risk, physical function etc) rather than just deterioration and mortality. Its calibration to clinical measures reflecting phenotypic aging often leads to divergent results, as it predicts broader health status rather than time-to-death *per se*. (Levine et al 2018)

These differences in training data and objectives result in each clock capturing a different aspect of aging biology. Thus, while PCGrimAge and DunedinPACE may produce more comparable results due to their focus on biological deterioration and mortality, PCPhenoAge can yield distinct outcomes because of its clinical biomarker basis and focus on overall healthspan. These differences highlight how the choice of clock depends on the research question—whether it's about predicting mortality, aging speed, or overall health status.

Literature

Lu AT, Quach A, Wilson JG, Reiner AP, Aviv A, Raj K, Hou L, Baccarelli AA, Li Y, Stewart JD, Whitsel EA, Assimes TL, Ferrucci L, Horvath S. DNA methylation GrimAge strongly predicts lifespan and healthspan. *Aging* (Albany NY). 2019 Jan 21;11(2):303-327. doi: 10.18632/aging.101684. PMID: 30669119; PMCID: PMC6366976.

Daniel W Belsky, Avshalom Caspi, David L Corcoran, Karen Sugden, Richie Poulton, Louise Arseneault, Andrea Baccarelli, Kartik Chamarti, Xu Gao, Eilis Hannon, Hona Lee Harrington, Renate Houts, Meeraj Kothari, Dayoon Kwon, Jonathan Mill, Joel Schwartz, Pantel Vokonas, Cuicui Wang, Benjamin S Williams, Terrie E Moffitt (2022) DunedinPACE, a DNA methylation biomarker of the pace of aging *eLife* 11:e73420

Levine ME, Lu AT, Quach A, Chen BH, Assimes TL, Bandinelli S, Hou L, Baccarelli AA, Stewart JD, Li Y, Whitsel EA, Wilson JG, Reiner AP, Aviv A, Lohman K, Liu Y, Ferrucci L, Horvath S. An epigenetic biomarker of aging for lifespan and healthspan. *Aging* (Albany NY). 2018 Apr 18;10(4):573-591. doi: 10.18632/aging.101414. PMID: 29676998; PMCID: PMC5940111.

R3.29: Overall, there is no balanced discussion of the science, methods, results and no mention of any limitations or other considerations.

We have now expanded the discussion of our results, and in particular of its links to previous research, and limitations in the following lines:

Line 221: *Kresovich et al.²⁰ found pregnancy to be linked with accelerated aging in Hannum, Horvath and PhenoAge clocks in a sample of 2,356 women from USA and Puerto Rico, yet these*

results attenuated after lifestyle-adjustment. We however found accelerated aging also of the nulliparous women in addition to the women with multiple pregnancies. We suggest these differences might result from Kresovich et al.²⁰ modelling parity ordinally per live birth, and thus it might be that potential age acceleration of the nulliparous women in their study was masked by the accelerated aging of the high-parity groups. Further, in their sample, parous women were significantly chronologically older than nulliparous women (53.7 years vs 57.0 years), which might partially explain why they did not find accelerated epigenetic aging for the nulliparous women. Ryan et al. found in 2018²¹ that each additional pregnancy was associated with accelerated epigenetic aging using the Horvath clock in 397 women aged 20-22 years from the Philippines, and no age acceleration of the nulliparous. In 2024, in the same population, Ryan et al.³ found that parous women were epigenetically older than nulliparous using Hannum, Horvath, DunedinPACE and PhenoAge clocks, but this effect was not significant for GrimAge. Again, while the result of accelerated aging in the high-parity groups aligns with our results, this contrasts our finding on the accelerated aging of also the nulliparous women. We suggest the discrepancies between ours and the studies by Ryan et al.^{3,21} may result from the fact that Ryan et al. found significantly accelerated aging mainly in epigenetic clocks that are trained to predict chronological age rather than mortality, and their study population was chronologically substantially younger at blood sampling than our sample (~22y. vs ~63.9y.). This highlights the dynamic nature of epigenetic patterns, and the differences in the biological aspects that different epigenetic clocks measure.

Line 244: Interestingly, women who reproduce early also have increased mortality risk and accelerated aging according to PCGrimAge and DunedinPACE, although the difference between early mothers on the one hand, and other groups on the other, did not remain significant in the adjusted models. This result is in line with previous studies that have observed a negative correlation between age at first reproduction, and either mortality^{5,10,11,12} or epigenetic aging²⁰.

Line 267: Our results on the accelerated aging in the high lifetime reproductive group are also in line with previous studies that have observed an increased number of children to associate with accelerated epigenetic aging^{3,20,21}.

Line 306: Limitations

Our study has some limitations. Firstly, we rely on self-reported smoking status, BMI, alcohol usage and information on the number and timing of childbirths. However, we have validated the information on childbirth for women born since the year 1950 (over 40% of women in our study sample) from the the Digital and Population Data Services Agency of Finland, and self-reported anthropometric measures have been previously shown to correspond accurately to clinically measured information⁴⁰ (see Supplementary Methods). Secondly, the wide range of birth years in our sample (1880-1957 for survival, 1909-1957 for EAA) may influence the estimation of survival

and aging patterns. We addressed this by controlling for cohort entry differences through left truncation in survival analyses and by including historical birth cohort as a covariate. The narrower birth year range in the epigenetic age acceleration (EAA) analyses (1909–1957) reduces this concern, with further adjustments using residual-based age acceleration measures and chronological age as a covariate. Furthermore, our study uses a twin design, which mitigates the challenges posed by the wide birth year range by controlling for genetic factors and shared early-life environments. This approach reduces biases from cohort differences and variations in early-life conditions, which are known to significantly influence later-life health and well-being⁴¹. Nonetheless, future research could benefit from more focused cohort designs with narrower age ranges to refine these analyses and minimize potential biases associated with generational differences. However, including participants across a broad temporal span has also allowed us to assemble a large study sample, enhancing statistical power and enabling robust examination of reproduction, aging, and lifespan.

Line 327: Conclusions

To our knowledge this is the first study to report distinct reproductive trajectories of pregnancy history in Western women. Our latent class analysis approach allowed us to capture the multidimensional nature of reproduction. Consistently with life-history theory predictions, women belonging to the classes of nulliparity and extreme lifetime reproductive output had increased mortality risk. For the first time, we were also able to link these differences in reproductive history and survival to biological aging, years before death. Our results confirm and extend prior findings of accelerated aging in both high-parity groups, and early childbearing groups, while also highlighting the dynamic nature of epigenetic aging processes across the lifespan. Furthermore, our twin design allowed us to control for family background and genetic effects on lifespan and aging. Together, these findings underscore the importance of considering both the timing and frequency of reproduction, as well as the choice of epigenetic aging markers, to better understand the trade-offs between reproduction, aging, and survival.

Reviewer #4 (Remarks to the Author):

The study sets out to test whether there is an association between reproduction patterns (number and timing of children) and survival in 17,080 Finnish twin women and further in a subset of 1,117 twins to test whether there is an association between reproductive patterns and the PCGrimAge epigenetic clock aging. The twins were born between 1880 and 1958 and data on reproduction was collected in 1975 when the twins were between 17 and 95 years old and followed through 2020. The period of the blood collection was 1994-2020 and the age range of the participants at blood drawing was 36-87 years.

R4.1: The paper has a very bold title: “Human reproduction comes at the expense of faster aging and a shorter life”, indicated that the associations demonstrated are causal. Furthermore, the study is mainly a confirmation of the well-known U-shaped association between parity and mortality risk. The title could therefore be expanded to include the sentence: “...and the absence of human reproduction also comes at the expense of faster aging and a shorter life”. The study also demonstrates that earlier timing of childbirth is associated with a negative aging profile. In other words, it is the large middle group (both in terms of number and timing) that have the most favorable profile in terms of survival and epigenetic clock, which is to be expected.

We agree that our results are associations rather than evidence for a causative process. In this study, we have controlled for genetic as well as lifestyle-related factors influencing reproduction and later health, and thus feel more confident that our results are not a mere reflection of the correlation between lifestyle attributes and underlying genes / early life parameters that would simultaneously influence reproduction, survival and aging. Furthermore, the directionality of this association is supported by Ryan et al. (2018), who found an association between increasing parity and accelerated biological aging: the authors then examined whether biological aging at baseline predicted the number of pregnancies in the coming four years, and found no evidence for reverse causation. Yet, we acknowledge that our results cannot be interpreted in a causative manner. We would still like to reflect in our title that the main finding discusses evolutionary theories of aging, namely the reproduction-aging trade-off of resource investment, while the association between nulliparity and mortality has been known for a long time, which likely reflects confounding by initial health status (eg. Grundy & Tomassini 2005).

We have now changed the title to “*From Cradle to Grave: Evidence for a Reproduction-Aging Trade-Off*” to indicate a less causative process between reproduction and later health, and emphasize our focus on the evolutionary trade-off.

Literature:

Ryan, C. P. *et al.* Reproduction predicts shorter telomeres and epigenetic age acceleration among young adult women. *Sci. Rep.* **8**, 11100 (2018).

Grundy, E. & Tomassini, C. Fertility history and health in later life: a record linkage study in England and Wales. *Soc. Sci. Med.* **61**, 217–228 (2005).

R4.2: Although the data and the analyses are of high quality, a major concern is that the twins included in the study are born between 1880 and 1958 – a nearly 80-year-long period. This makes the interpretation of the study very difficult, also in terms of the confounding control. For example, it is probably a very different covariate to have seven years of lifetime education for a woman giving birth in the year 1900 compared to a women giving birth in the year 1980. It is not clear that the historical birth cohort variable is able to accounts for such challenges.

Thank you for your thoughtful comment. We fully acknowledge that including twins born over such a long period—nearly 80 years—introduces complexities in interpreting the study results. As you pointed out, covariates like lifetime education could indeed have very different implications for individuals born in different historical periods, and accounting for these differences is not straightforward.

However, fully accounting for all such temporal variations would require including a vast number of additional parameters and multi-way interactions, which would drastically increase the complexity of the model and potentially introduce overfitting or loss of statistical power. While we have taken steps to control for historical cohort effects through the inclusion of birth cohort as a variable, we recognize that a more granular approach could provide additional insights, but it would come at the cost of significantly complicating the analysis.

In light of these constraints, we maintain that our approach, which includes birth cohort as a key variable, provides a reasonable balance between capturing historical context and avoiding an unmanageable model. We also adjust for key lifestyle factors, including smoking, alcohol use, BMI, and education, which are well-established confounders in studies on aging and health. Additionally, our use of a twin design allows us to control for genetic factors and shared early-life environments, thus mitigating some of the challenges posed by the large time span.

In summary, while we acknowledge the challenges introduced by the long time frame of the study cohort, we believe that our approach, while not exhaustive in addressing every potential confounder, offers a robust and interpretable model. The inclusion of historical birth cohort, along with the controls for key lifestyle factors and the use of a twin design, provides a strong foundation for understanding the relationships between parity and later health.

We have now added the description of this limitation in our manuscript line 312: *“Secondly, the wide range of birth years in our sample (1880-1957 for survival, 1909-1957 for EAA) may influence the estimation of survival and aging patterns. We addressed this by controlling for cohort entry differences through left truncation in survival analyses and by including historical birth cohort as a covariate. The narrower birth year range in the epigenetic age acceleration (EAA) analyses (1909–1957) reduces this concern, with further adjustments using residual-based age acceleration measures and chronological age as a covariate. Furthermore, our study uses a twin design, which mitigates the challenges posed by the wide birth year range by controlling for*

genetic factors and shared early-life environments. This approach reduces biases from cohort differences and variations in early-life conditions, which are known to significantly influence later-life health and well-being⁴¹. Nonetheless, future research could benefit from more focused cohort designs with narrower age ranges to refine these analyses and minimize potential biases associated with generational differences. However, including participants across a broad temporal span has also allowed us to assemble a large study sample, enhancing statistical power and enabling robust examination of reproduction, aging, and lifespan.”

R4.3: Furthermore, it is remarkable that more than a third of the total sample of women do not have any children. This can be explained by the fact that the study population is between 17 and 95 years old, so the “no children group” consists of a combination of young women/girls who have not started trying or succeeded in getting children and a group of mainly middle-aged and elderly who have not had any children. Whether it is meaningful to combine these groups into one is highly questionable.

We thank the reviewer for pointing out the suspiciously high number of childless women in our data (N=6,392/17,080, 37.42%). Thanks to this, we started investigating whether we have any additional information on the reproductive history of these women. We found a separate, archive dataset dating from the data collection period and described in Kaprio et al. (1987) that specifies the number of children for the cohort participants: in that dataset, there were 2,244 women who appeared childless in our original data, but had at least one child - all of those are women born between 1939 and 1949. We are unsure how this data became corrupted in our general dataset: we have now removed all data on these 2,244 women, leaving us with a total sample size of 14,836 for the survival analysis, and 1,054 women for the epigenetic analyses. The total number of childless women in our data is now 4,148 (27.78% of the total 14,931). We have now rerun all of our analyses with this reduced sample size, and corrected the sample size in the Manuscript, Supplementary Methods and Extended Data. All of our answers in this response letter that include model comparison have also been conducted using this dataset of 14,931 women. **We heartily thank the reviewer for identifying and pointing out this flaw in our data - we underline at the same time that this issue does not affect our inference of reproductive trajectories (LCA), which only included women with documented live births, and for which all data was manually verified.** Only the group of women added as “nulliparous” for the later stages of the analysis was reduced.

Still, the proportion of childless in this current data is ~28%, which is higher than 21%, that has been previously reported for the women born 1950-1957 in our cohort after 33 years of follow-up (Rose et al. 2022). Thus, the “no children” group indeed likely consists partially of women who may not have started their reproductive career yet at the time of the questionnaire in 1975. However, the live birth data (or absence thereof) has been validated for all participants born after 1949 from the Digital and Population Data Services Agency Finland. Thus, the youngest women for whom only questionnaire information was used were 26 years old, not 17. We would also like

to note that in our latent class solution, the number of children was coded as missing and not 0 for each participant after the age at which they responded to the 1975 questionnaire. We would like to ask the reviewer to see our answer to comment R3.11 regarding the age of the participants at the assessment of childbirth data.

Kaprio, J. *et al.* Design and Sampling Considerations, Response Rates, and Representativeness in a Finnish Twin Family Study. *Acta Genet Med Gemellol.* 36: 79-93 (1987).

Rose, R.J. *et al.* Alcohol consumption at age 18-25 and number of children at a 33-year follow-up: Individual and within-pair analyses of Finnish twins. *Alcohol Clin Exp Res.* Aug;46(8):1552-1564. (2022).

R4.4: Furthermore, the latent class model including six classes defines both the very big group of women with no children and the very small group of women with a high average number of children (6.7). The remaining groups, which consist of the majority of participants, all are within a band from 2.0 to 2.4 children with the timing being the main difference. So, basically, all the analyses hinge on the two extreme groups that form the basis for the very bold claims in the title. The middle groups (maybe with the exception of class 1 of very early births) show very little differences in survival and aging profile when confounders are adjusted for. As the no children group is a mixed group, and the class 6 is a small group with a very high number of children, the interpretation of the association between human reproduction and aging/survival is very difficult regarding the “typical” range of human reproduction. On top of this, there are difficulties in the interpretation of confounder control over the nearly eight decades of birth cohorts included.

We understand the reviewer’s concerns regarding the interpretation of our results. As we mention above in comment R4.3, the no children group consists mainly of women whose data has been validated from the Digital and Population Data Services Agency and women who had already completed their reproductive history by the time of the 1975 questionnaire. However there might still be some women who had children after 1975 , and who were born before 1950 (i.e. who were at age of ≥ 26 yrs in 1975) and thus we might have incomplete information of their reproductive history, and we acknowledge this might be a limitation to our study.

We agree with the reviewer that our results suggest that in this population, the optimal number of offspring, from the perspective of maternal health, is between 2 and 2.4 children (classes 2-5), with no major differences observed between the middle groups. The average number of children in our sample was 2.4 (see Supplementary Results). Similarly, the optimal timing of reproduction (classes 2-5) aligns with the average age of reproduction in our sample (overall average 27.3, with 24.4 and 29.8 for first and last childbirth, see Supplementary Results). We have now added a paragraph to our discussion addressing this notation in line 276:

Typical reproduction is the optimal reproduction

Our results indicate that regarding later-life maternal health, the best number of offspring seems to be between 2 and 2.4 children (classes 2-5), while the average number of children in our sample was 2.4 (see Supplementary results). In parallel, the best timing of reproduction (classes 2-5) was also around the average time of reproduction in our sample (overall average 27.3, with 24.4 and 29.8 for first and last childbirth, see Supplementary Results). This convergence of “optimal” and average reproductive behaviour is not surprising under the life-history theory: however, we underline that it is likely driven both by biological parameters, and by socio-economic and cultural constraints valid in this particular context (and reflecting e.g. parental resources, housing, healthcare, childcare, and associated strain on parents in general and mothers in particular). Thus, it is important to note that these patterns reflect the characteristics of our specific study population and may not generalize directly to other cultural or demographic contexts. Furthermore, reproductive decisions are likely influenced by a range of individual and environmental factors, with trajectories that may be adaptive for one person not necessarily benefiting another. Thus, while our results align with the idea that the fertility patterns are shaped by natural selection to balance offspring quantity and somatic self-maintenance³⁹ within the context of a given environment, caution is needed when interpreting these findings more generally. Further research is needed to understand how these findings apply across diverse human populations.

Our latent class approach allowed us to show that within the typical reproductive pattern range (classes 2-5), the differences in survival and aging profiles are minimal after adjusting for confounders. This suggests that reproductive timing and number of offspring may have a smaller impact on aging and survival than we initially expected. This result is supported when repeating our analysis using the number of births and age at first child as predictors of survival and epigenetic age acceleration (see Extended Figure 7), reinforcing our confidence that we have not overlooked relevant signals, particularly concerning the timing of childbirth.

REVIEWER COMMENTS:

Reviewer #1 (Remarks to the Author):

The revision has addressed the concerns raised by the referees to the extent that the structure of the data set allows, and the limitations of the data set and analysis are now more clearly communicated.

Reviewer #2 (Remarks to the Author):

The authors have mostly addressed my comments. In one case, i am concerned about the statistical approach. In their response to review, they present their random effects model as accounting for factors shared by siblings within a family. This is incorrect. In fact, any family-level factor that contributes to the outcome that is not already explained by the included regressors will BIAS the effect estimates of interest. We should be particularly concerned about such bias in the case of this paper because many factors shared by siblings may influence both fertility and biological aging. Instead of using a random effects model, the authors should be using what economists refer to as a fixed effects model (sometimes called a within-group regression; see here <https://www.sciencedirect.com/topics/social-sciences/fixed-effects-model>) that focuses the hypothesis test on differences between siblings in a family. There are a few ways to implement this. But the easiest way to think about it is as a "Sibling Difference" design where sibling differences in biological age are regressed on sibling differences in fertility.

We thank the reviewer for bringing this issue to our attention - indeed this needed to be clarified better in our text. We agree with the reviewer that random-effect models indeed impose a “restricted” distribution of shared family-level intercepts, and that this can under-estimate or in extreme cases bias estimates.

A within-pair-difference model, as suggested here, would indeed be able to do that - although it would raise other questions about covariate formulation, as the modelling unit becomes the twin pair and not the individual, and consequently we need to account for both average value and pair-difference in each included covariate (the latter having no standard definition for e.g. study level, quantitative tobacco use, etc). And most importantly, it would strongly reduce our sample size, as we only focus on women here, and we would not be able to include the female co-twin of opposite-sex dizygotic pairs in our model anymore.

Our approach using family level random effect is thus a compromise that aims at correcting for family relatedness, where the introduction of the random effect is only meant to avoid violating the assumption of independence of observations that we would encounter when using a standard linear regression. Our approach has been described with regards to twin models by Fasola et al. (2021) as generalized linear mixed models (GLMMs), where the authors have also compared the approach to other methods used in twin analysis, such as the within-pair model. GLMMs are commonly used in twin data analyses, as they allow a response variable to be explained by both fixed effects (which apply uniformly to all individuals) and random effects (which vary by family, capturing sources of similarity between twins), enhancing the robustness of statistical inferences by appropriately modeling the dependencies inherent in twin data. This method has been widely used in twin studies (Yuan et al. 2014, Fasola et al. 2021), also in our cohort (Huang et al. 2020, Gustavson et al. 2025).

However, we understand that describing the analyses as "twin model analyses" in the manuscript may be misleading and we agree that there was a lack of clarity about how the models and their results should be interpreted, which we have now corrected in the manuscript. We now clearly state that the family ID is only meant to correct for statistical non-independence between the twins from the same family throughout the manuscript. In addition, we have made sure that there is no confusion regarding potential corrections for shared factors between co-twins, as our models do not allow this.

Fasola, S.; Montalbano, L.; Cilluffo, G.; Cuer, B.; Malizia, V.; Ferrante, G.; Annesi-Maesano, I.; La Grutta, S. A Critical Review of Statistical Methods for Twin Studies Relating Exposure to Early Life Health Conditions. *Int. J. Environ. Res. Public Health* 2021, 18, 12696. <https://doi.org/10.3390/ijerph182312696>

Gustavson, D. E., Bell, T. R., Buchholz, E. J., Zellers, S., Luczak, S. E., Reynolds, C. A., Finch, B. K., Nygaard, M., Catts, V. S., Christensen, K., Finkel, D., Kremen, W. S., Latvala, A., Martin, N. G., McGue, M., Mewton, L., Mosing, M. A., Panizzon, M. S., Plassman, B. L., ... Franz, C. E. (2025). Genetic and environmental influences on alcohol consumption in middle to late life. *Psychology of Addictive Behaviors*. <https://doi.org/10.1037/adb0001052>

Huang, Y., Ollikainen, M., Muniandy, M., Zhang, T., van Dongen, J., Hao, G., van der Most, P. J., & Wang, X. (2020). Identification, heritability, and relation with gene expression of novel DNA methylation loci for blood pressure. *Hypertension*, 76(1). <https://doi.org/10.1161/HYPERTENSIONAHA.120.14973>

Yuan, W., Xia, Y., Bell, C. et al. An integrated epigenomic analysis for type 2 diabetes susceptibility loci in monozygotic twins. Nat Commun 5, 5719 (2014). <https://doi.org/10.1038/ncomms6719>

Reviewer #3 (Remarks to the Author):

The authors have voluminously responded to previous reviewer comments. It is strongly suggested to just put these limitations into the manuscript rather than inching up to them, writing 4 paragraphs in response, and then disregarding in the text.

Some points for perspective:

R3.1 Pregnancy is a well-known physiological stressor. As noted by multiple reviewers, this isn't a new finding or surprising. Well documented phenomena across populations.

We agree that pregnancy, as well as the associated burden of raising children are well-known physiological stress factors, which was the basis for initiating this study.

R3.2 The sub-sample of epigenetic measures taken between 1994-2020, along with the table 1 data on the underlying population could be a case-study for selection bias, and plausible avenues for why what is observed in the data is observed. Based upon the voluminous responses to the reviewers, consulting a scientist who understands this concept would have been a better use of everyone's time. In short, there appears to be survival based requirement to be included, along with a response requirement. The reader doesn't know the baseline denominator of potential subjects for the baseline assessment. The epigenetic data was a convenience sample and not representative of the overall cohort.

In the Finnish Twin Cohorts the high response rates in follow-ups since the founding of the study in 1975 indicate the FTC is broadly representative of the general population (Kaprio et al. 2019). Long-term analyses show no significant differences from the general population in mortality, cancer incidence, or genetic studies (Kaprio 2013, Skytthe et al, 2019). For our study, we utilized all available DNA methylation data among the cohort participants collected from various targeted studies and general expansions without phenotype-based selection. Studies performed utilising epigenetic clocks in the cohort data show very consistent results with other population studies. However, we acknowledge that there might still be some selection bias related to the prerequisite to be alive and answer the questionnaire in 1975 and that some epigenetic data has been gathered in purpose to target specific traits, and we have now added this to the Limitations-section of our Main text:

Third, while the Finnish Twin Cohort has high response rates and has been shown to be broadly representative of the general population^{7,42,43}, we acknowledge that some selection bias may be present due to the requirement for participants to be alive and respond to the baseline survey in 1975. Additionally, the available DNA methylation data was collected from various targeted

studies and general expansions, meaning that some samples were selected to investigate specific traits rather than being fully representative of the overall cohort. Nevertheless, studies utilizing epigenetic clocks within this dataset have produced results that align well with other population studies⁴⁴⁻⁴⁶, suggesting that major selection effects are unlikely.

7. Kaprio, J. et al. The Older Finnish Twin Cohort - 45 Years of Follow-up. *Twin Res. Hum. Genet. Off. J. Int. Soc. Twin Stud.* 22, 240–254 (2019)

42. Kaprio J. The Finnish Twin Cohort Study: An Update. *Twin Res. Hum. Genet.* 16, 157-162 (2013).

43. Skytthe A, Harris JR, Czene K, et al. Cancer Incidence and Mortality in 260,000 Nordic Twins With 30,000 Prospective Cancers. *Twin Res. Hum. Genet.* 22, 99-107 (2019).

44. Lundgren, S. et al. BMI is positively associated with accelerated epigenetic aging in twin pairs discordant for body mass index. *J. Intern. Med.* 292, 627–640 (2022).

45. Kankaanpää A, Marttila S, Törmäkangas T, et al. The role of adolescent lifestyle habits in biological aging: A prospective twin study. *Elife* 11, e80729 (2022).

46. Föhr, T. et al. Metabolic syndrome and epigenetic aging: a twin study. *Int. J. Obes. (Lond.)* 48, 778–787 (2024).

R3.3 While the goal of the research may indeed have a causal lens, the methods applied are not estimating a causal effect, but the language in the text throughout implies causality. This was not changed, despite reviewer comments.

We have now clarified our manuscript to indicate that our findings are associations, and that any discussion about causality is speculative. We have also added this point to our limitations-section:

Finally, while our study examines associations between reproductive history, epigenetic aging, and survival, the observational nature of our data and the methods used do not allow for causal inference. Although we employ rigorous statistical approaches to control for potential confounders, our findings should be interpreted as associations rather than causal effects. Future research using for example, quasi-experimental designs, such as Mendelian randomization or other instrumental variable approaches, may help to better establish causal relationships.

R3.4 The Cox regression survival analysis does not define how follow up was calculated or what time 0 was, making it impossible to interpret. It takes a puzzle master to figure out when the various covariates were measured/included as well, and still unclear how they align.

We have now specified the start of the follow-up and the adjustment for delayed entry throughout the manuscript, and described our method in the Main text and the Supplementary methods (see below). We have also specified the information on when the covariates were collected for the survival and epigenetic aging analyses.

In the Main text: Methods line 537:

The start of the follow-up was birth, and we used left truncation to account for delayed entry from the time of the first questionnaire in 1975.

Supplementary methods: 3.2 Survival:

The follow-up started at birth, and as the study participants only entered the study provided they answered the 1975 questionnaire, we used 1975 as an entry date for left truncation in our model.

R3.5 BMI is not a lifestyle factor.

We have now changed the description of our “lifestyle” adjustments to “risk factors” throughout the manuscript.

R3.5 Yes, the application of latent class modeling to derive “distinct” classes helps statistically, but what any of the classes represent beyond the nulliparous and those with a high number of pregnancies. Despite this, there’s nothing really going on except on the tails of the distribution, and despite nulliparous women being at the highest risk compared to whatever latent class 3 is, the text largely ignores this and spends most of the time speculating and interweaving supported theories with evolutionary reproduction factors and focusing the conclusions on this.

We appreciate the reviewer’s comment and agree that the most pronounced differences in survival and aging outcomes appear at the extremes of the reproductive distribution, particularly for nulliparous women and those with a high number of pregnancies. However, our latent class modeling approach was designed to explore whether there were additional meaningful distinctions in survival and aging patterns beyond these well-known extremes. Since we could not assume in advance what the result would be, our goal was to identify potential heterogeneity within the typical reproductive patterns (classes 2–5) rather than focusing solely on predefined groups.

Our findings indicate that, after adjusting for confounders, differences within these middle-range classes were minimal. While this outcome may suggest that reproductive timing and moderate variations in offspring number are linked to smaller variations with our outcomes than initially expected, our approach was necessary to establish this empirically. To further support this conclusion, we also conducted analyses using the number of births and age at first child as continuous predictors (see Extended Data Figure 7 and Supplementary Results), reinforcing our confidence that no major overlooked patterns exist, particularly concerning childbirth timing. We

have described this rationale in our Main text under the subheading: *Typical reproduction is the optimal reproduction.*

Reviewer #4 (Remarks to the Author):

The authors have been very responsive to the reviews and revised the manuscript satisfactorily.

Reviewer #2 (Remarks to the Author):

The authors have addressed my comment. I note that the approach the authors opt for is vulnerable to bias in the case that unmeasured factors are correlated with both the outcome and the unit of the random effect (in this case, the family unit). Such factors are almost certainly present -- genetics, unmeasured variability in socioeconomic circumstances, family norms in health behaviors, etc. There is a formal test to determine if random effects are valid in this setting proposed by Houseman (e.g. see here <https://www.sciencedirect.com/topics/social-sciences/hausman-test>). It almost always rejects the random effects approach, though. An alternative might be to evaluate the consistency of results using alternative methods to account for the non-independence of observations within families: clustering standard errors at the family level, generalized estimating equation (GEE) models. At minimum, the authors should acknowledge the vulnerability of their analysis to this type of confounding.

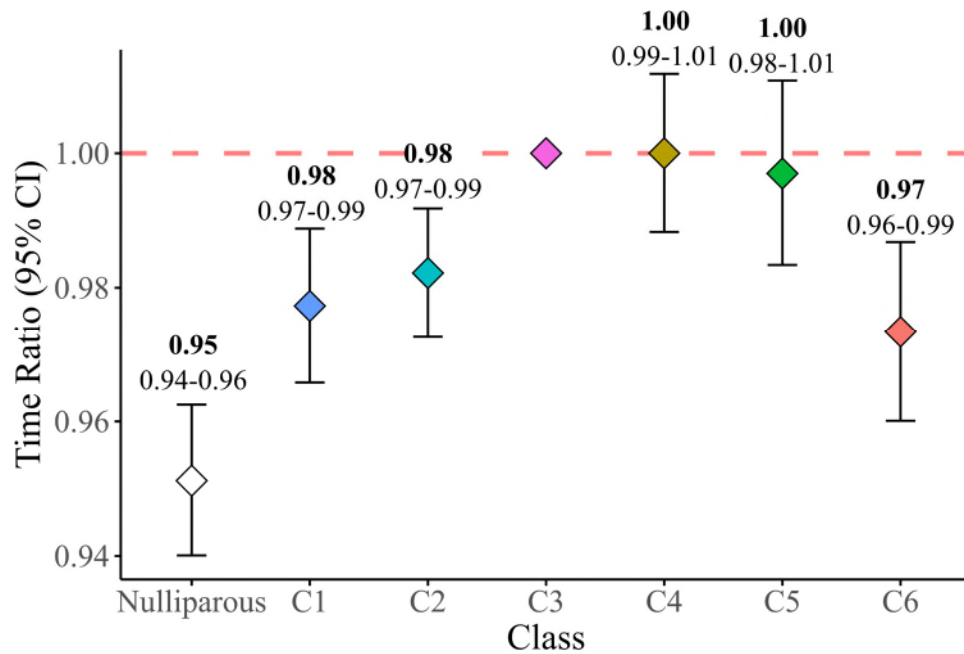
Response:

We thank the reviewer for raising the important issue of unobserved heterogeneity—specifically, the assumption in mixed-effects models that the random effect (in our case family-level intercept) is not correlated with the predictors (such as reproductive class or covariates). We agree that if this assumption is violated, it could bias the model estimates and their interpretation.

In response to this concern, we conducted several additional analyses to assess the robustness of our model and the potential for bias. These results are developed below and have been added to the Main text and Supplementary Methods.

Due to the large size of our dataset (~9,500 families), the fixed-effect model necessary for the Hausman test would not converge on the Finnish IT center for science computing grid. As a result, we proceeded directly to testing generalized estimating equation (GEE) models as suggested by the reviewer. As we are not aware of any GEE implementation for Cox proportional hazards models matching our data structure (left-truncated, right-censored data, proportional weighting by latent class and time-to-event outcomes), we employed Accelerated Failure Time (AFT) models using the GEE framework, which allow modeling of survival time (rather than hazard ratios) while adjusting for right censoring, latent class weights, and covariates. However, the only currently available implementation in R to our knowledge (`aftgee`) is unable to account for left truncation (i.e. the fact that participants entered the study in 1975 at different ages). The results of the AFT-GEE model adjusted only for historical cohorts supports our original conclusions (see results below). In this “unadjusted” model, the nulliparous women have a time ratio of 0.95 (95% CI: 0.94–0.96), indicating a 5% shorter survival time, while class 1 and 2 have time ratios of 0.98 (CI: 0.97–0.99), corresponding to 2% reduction in survival time. Finally, class 6 has a time ratio of

0.97 (CI: 0.96–0.99), corresponding to a 3% reduction in lifespan. Unfortunately, as in the case of the Hausman test, the risk factor (alcohol use, smoking, BMI and education) adjusted model did not converge in reasonable time (10 days) on our computing grid, despite repeated attempts. Because of this difficulty, and of AFT-GEE models' inability to account for left truncation, we report this modeling strategy here for completeness but do not include it in the main manuscript.



Unadjusted survival model in generalized estimating equation -framework. Time ratios with 95% confidence intervals for each latent class from an accelerated failure time model weighted with class-specific posterior probabilities, and adjusted for historical cohort (N=14,836). The model accounts for family ID as a clustering variable with exchangeable correlation structure. The red dashed line indicates the reference class in the model (class 3).

As suggested by the reviewer, we also re-estimated our survival models without random effects but with cluster (here family) robust standard errors (Extended Data Table 5, reproduced below). The survival model estimates with cluster robust standard errors are consistent with our original results, providing additional evidence for the robustness of our findings. For epigenetic age acceleration analyses, the models presented in the manuscript already use cluster robust standard errors Bolck–Croon–Hagenaars method with type=COMPLEX in Mplus).

Extended Data Table 5. Survival models with cluster robust standard errors. “Original manuscript model” refers to the Cox proportional hazards model weighted with class-specific posterior probabilities accounting for left truncation with Family ID as a random effect. “Cluster-robust SE”-models refer to the same models without random effects, but where family-level clustered standard errors have been used. Sample sizes (N), Hazard ratios (HR) and 95% confidence intervals (LL95%-UL95%) are given for each reproductive class from each model. Model A adjusts for historical cohort, while models B and C additionally adjust for risk factors (alcohol use, smoking, BMI and education) with imputed data and non-imputed data, respectively.

A. Survival Unadjusted model for the complete case sample (N=14836)						
Class	Original manuscript model			Cluster-robust SE		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.4273	1.2935	1.575	1.3619	1.2596	1.4725
1	1.1835	1.051	1.3326	1.1641	1.0695	1.2671
2	1.1049	0.992	1.2307	1.093	1.0122	1.1802
4	1.0182	0.9008	1.151	1.0083	0.9242	1.1001
5	1.014	0.8915	1.1533	1.0142	0.9186	1.1197
6	1.2479	1.0657	1.4613	1.2498	1.1208	1.3936

B. Survival Risk factor adjusted model using imputed data (N=14836)						
Class	Original manuscript model			Cluster-robust SE		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.4147	1.282	1.5612	1.3594	1.252	1.4761
1	1.0994	0.9756	1.239	1.0847	0.9926	1.1853
2	1.0738	0.9639	1.1963	1.069	0.9865	1.1584
4	1.0279	0.9093	1.1619	1.0273	0.9396	1.1231
5	1.0203	0.8969	1.1606	1.0228	0.9227	1.1337
6	1.1649	0.9942	1.3648	1.1651	1.0375	1.3084

C. Survival Risk factor adjusted model for the complete case sample (N=10975)						
Class	Original manuscript model			Cluster-robust SE		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.3336	1.2034	1.4779	1.2944	1.1877	1.4107
1	1.0679	0.9413	1.2116	1.0577	0.9633	1.1613
2	1.0619	0.9486	1.1886	1.0611	0.9767	1.1529
4	1.0418	0.9179	1.1824	1.0396	0.9467	1.1417
5	1.0187	0.8917	1.1638	1.0202	0.916	1.1362
6	1.1611	0.9884	1.364	1.1589	1.0292	1.3048

Here, we propose a third and more straightforward method to test for the effect of unmeasured family-level effects on our results: as a control, we simply down-sampled our dataset to a single individual per family, removing the twin structure entirely from our data. This “singleton” approach was repeated 100 times, each time randomly selecting one individual per family. For each random singleton subset, we modeled our left-truncation adjusted Cox proportional hazard models weighted with class-specific posterior probabilities in R, as well as our epigenetic age acceleration (EAA) models using the Bock-Croon-Hagenaars method in Mplus. We summarized the hazard ratios, estimates and 95% confidence intervals across these models using the function `rma(method="FE")` from the R package `metafor()`. Comparisons of hazard ratios (survival) and estimates (EAA) along with their confidence intervals from these singleton models to our original models are presented in Extended Data Table 6, and the estimates from the individual models are shown in Extended Data Figure 8 (reproduced below). Sample sizes vary slightly in non-imputed Model 2 due to differences in the number of individuals with missing data across the random subsets. In contrast, sample sizes remain consistent for Survival Model 1 (complete-case model which includes no missing data) and for the imputed Model 2 (where missing values are replaced by imputed values). Sample sizes remain consistent in the non-imputed BCH Model 2 because it employs Full Information Maximum Likelihood (FIML), which retains all observations by using available data without excluding cases due to missing covariate values. These singleton analyses unambiguously support our original conclusions regarding the between-class differences in both survival and EAA.

Extended Data Table 6. Results from 100 repeated “Singleton”-models. One individual was randomly selected from each family, and results were aggregated using a fixed-effects meta-analysis across 100 models with different singleton subsets. A-C. Mortality hazard ratios with 95% confidence intervals for each latent class from a Cox proportional hazards model weighted with class-specific posterior probabilities. The model accounts for left truncation and adjusts for historical cohort (A), and additionally adjusts for risk factors (alcohol use, smoking, BMI and education) with imputed data (B), and non-imputed data (C). D-E. Estimates and 95% confidence intervals for epigenetic age acceleration (EAA) by the PCGrimAge clock for each class from a weighted Bolck-Croon-Hagenaars model (E), with additional adjustments for risk factors (E).

A. Survival Unadjusted survival model for the complete case sample						
Class	Original manuscript model (N=14,836)			Singleton model (N=9,496)		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.4273	1.2935	1.575	1.3869	1.3705	1.4035
1	1.1835	1.051	1.3326	1.2001	1.1827	1.2178
2	1.1049	0.992	1.2307	1.1098	1.095	1.1248
4	1.0182	0.9008	1.151	0.9938	0.9786	1.0091
5	1.014	0.8915	1.1533	1.0093	0.9935	1.0253
6	1.2479	1.0657	1.4613	1.2577	1.2336	1.2822

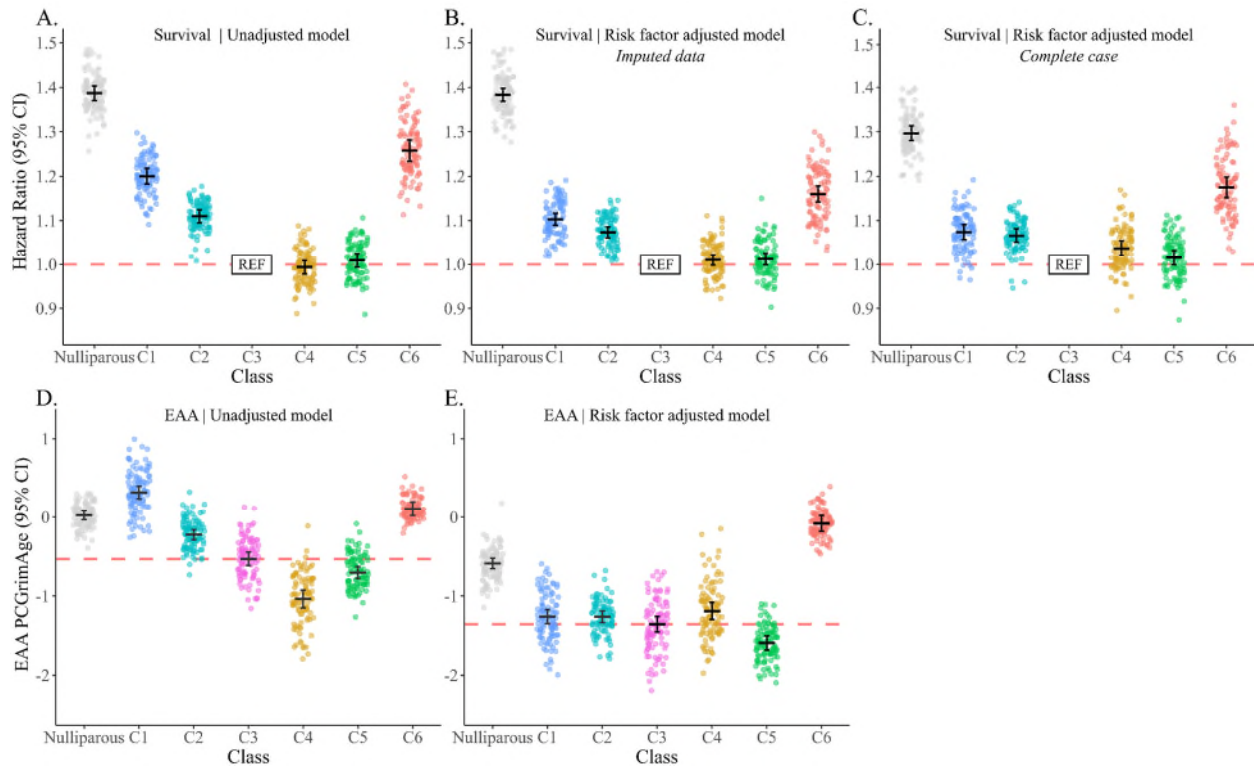
B. Survival Risk factor adjusted survival model using imputed data						
Class	Original manuscript model (N=14836)			Singleton model (N=9496)		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.4147	1.282	1.5612	1.3834	1.3688	1.3981
1	1.0994	0.9756	1.239	1.1032	1.0897	1.1169
2	1.0738	0.9639	1.1963	1.0737	1.0617	1.0859
4	1.0279	0.9093	1.1619	1.0106	0.9984	1.023
5	1.0203	0.8969	1.1606	1.0123	0.9988	1.0261
6	1.1649	0.9942	1.3648	1.1598	1.142	1.1778

C. Survival Risk factor adjusted survival model for the complete case sample						
Class	Original manuscript model (N=10975)			Singleton (N _{Mean} =6031.15, N _{Range} =6,011-6054)		
	HR	LL95%	UL95%	HR	LL95%	UL95%
0	1.3336	1.2034	1.4779	1.2966	1.2803	1.3131
1	1.0679	0.9413	1.2116	1.0744	1.0574	1.0917
2	1.0619	0.9486	1.1886	1.0663	1.0512	1.0817
4	1.0418	0.9179	1.1824	1.0373	1.0207	1.0541
5	1.0187	0.8917	1.1638	1.0158	0.9991	1.0328
6	1.1611	0.9884	1.364	1.1747	1.1516	1.1983

D. EAA Unadjusted PCGrimAge epigenetic age acceleration model						
Class	Original manuscript model (N=1054)			Singleton (N=611)		
	Estimate	LL95%	UL95%	Estimate	LL95%	UL95%
0	0.11	-0.35844	0.57844	0.0257	-0.03	0.0814
1	0.374	-0.26496	1.01296	0.32	0.2386	0.4015
2	-0.335	-0.80344	0.13344	-0.224	-0.2869	-0.161
3	-0.649	-1.29188	-0.00612	-0.5299	-0.6147	-0.4451
4	-1.045	-1.88976	-0.20024	-1.0356	-1.1467	-0.9246
5	-0.57	-1.24032	0.10032	-0.703	-0.7794	-0.6265
6	-0.076	-0.79924	0.64724	0.1086	0.021	0.1962

E. EAA Risk factor adjusted PCGrimAge epigenetic age acceleration model						
Class	Original manuscript model (N=1054)			Singleton (N=611)		
	Estimate	LL95%	UL95%	Estimate	LL95%	UL95%
0	-0.5	-1.03312	0.03312	-0.5865	-0.6546	-0.5184
1	-1.085	-1.75532	-0.41468	-1.2583	-1.3467	-1.1699
2	-1.312	-1.85884	-0.76516	-1.2611	-1.336	-1.1863

3	-1.461	-2.21952	-0.70248	-1.356	-1.4538	-1.2581
4	-1.164	-1.93036	-0.39764	-1.1887	-1.2971	-1.0804
5	-1.5	-2.24284	-0.75716	-1.5905	-1.6802	-1.5008
6	-0.146	-0.91236	0.62036	-0.0801	-0.1824	0.0222



Extended Data Figure 8. Results from 100 repeated “Singleton”-models. One person was randomly selected from each family, and the estimates (shown as individual points) were collected across all 100 models with different singleton subsets. The black lines with error bars represent the aggregated estimates with 95% confidence intervals across all 100 models calculated by a fixed-effects meta-analysis. A-C. Mortality hazard ratios (bolded) with 95% confidence intervals for each latent class from a Cox proportional hazards model weighted with class-specific posterior probabilities. The model accounts for left truncation and adjusts for historical cohort (A; N=9,496), and additionally adjusts for risk factors (alcohol use, smoking, BMI and education) with imputed data (B; N=9,496), and non-imputed data (C; N_{Mean}=6,031.15, N_{Range}=6,011-6,054). D-E. Estimates (bolded) and 95% confidence intervals for epigenetic age acceleration (EAA) by the PCGrimAge clock for each class from a weighted Bolck-Croon-Hagenaars model adjusted for chronological age (D; N=611), with additional adjustments for risk factors (E; N=611). The dashed red line is drawn for comparison to survival analysis at the estimates for class 3.

We have now incorporated these results to the manuscript as follows:

Main text:

Line 181: *We used alternative modeling strategies to assess the robustness of our results. First, we reanalyzed survival time using fixed-effects models with cluster-robust standard errors. Second, we implemented a “singleton” approach by drawing 100 random subsamples, each including one individual per family from the 9,496 families, and reran our survival models on these samples. Both approaches consistently showed elevated hazard ratios for nulliparous women, early mothers, and the high lifetime reproductive group (classes 0, 1 and 6). The increased risks for nulliparous women and the high lifetime reproductive group remained statistically significant after adjusting for risk factors, underscoring the robustness of these associations (Supplementary Methods, Extended Data Fig. 8, Extended Data Tables 5 and 6).*

Line 221: *These results were also supported by our “singleton” approach with 100 random draws of individuals from each of the 611 families in our epigenetic age acceleration analyses (Extended Data Table 6, Extended Data Fig. 8).*

Supplementary methods:

3.5 Alternative modeling strategies

To test whether our results would be impacted by the modeling framework, we modeled our original Cox proportional hazards models with cluster robust standard errors (Extended Data Table 5). The models were otherwise the same as our original models, but Family ID was used as a clustering variable instead of a random intercept. Finally, we down-sampled our dataset to a single individual per family, removing the twin structure entirely from our data. This “singleton” approach was repeated 100 times, each time randomly selecting one individual per family. For each random singleton subset, we modeled our left-truncation adjusted Cox proportional hazard models weighted with class-specific posterior probabilities in R, as well as our epigenetic age acceleration (EAA) models using the Bock-Croon-Hagenaars method in Mplus. We summarized the hazard ratios, estimates and 95% confidence intervals across these models in a fixed-effects meta-analysis using the R package metafor⁴⁴.

Reviewer #3 (Remarks to the Author):

The authors have largely responded to everything and updated the manuscript. One point remains which needs updating, R 3.2:

The following statement should be removed from the text: "Nevertheless, studies utilizing epigenetic clocks within this dataset have produced results that align well with other population studies⁴⁴⁻⁴⁶, suggesting that major selection effects are unlikely."

This is incorrect and not an approach or valid way to address or assuage the selection bias.

Response:

We thank the reviewer for noting this. We have now removed the sentence mentioned by the reviewer, and modified the limitations section at line 333 onwards:

"Additionally, the available DNA methylation data were drawn from a combination of targeted studies and broader cohort expansions. As a result, some samples were specifically selected to investigate particular traits, rather than to represent the full cohort. Consequently, analyses based on this subsample of epigenetic data should be interpreted with caution, as the participants included may be subject to selection bias and may not fully reflect the general population by their physiological condition."