Detecting telomere elongation in longitudinal datasets: Analysis of a proposal by Simons, Stulp and Nakagawa

Daniel Nettle[1] & Melissa Bateson

Centre for Behaviour and Evolution & Institute of Neuroscience, Newcastle University, Newcastle, UK

Abstract

Telomere shortening has emerged as an important biomarker of aging. Longitudinal studies consistently find that, although telomere length shortens over time on average, there is a subset of individuals for whom telomere length is observed to increase. This apparent lengthening could either be a genuine biological phenomenon, or simply due to measurement and sampling error. Simons, Stulp and Nakagawa [*Biogerontology* 15: 99-103, 2014] recently proposed a statistical test for detecting when the amount of apparent lengthening in a dataset exceeds that which should be expected due to error, and thus indicating that genuine elongation may be operative in some individuals. The test is however based on a restrictive assumption, namely that each individual's true rate of telomere change is constant over time. This assumption is unrealistic, since stress and life events are thought to affect telomere dynamics, and such events occur episodically. Here we show, using simulated data that realistically mirrors empirically-observed telomere parameters, that when the assumption of a constant true rate for each individual does not hold, the test fails to detect true lengthening in a wide range of cases where it does exist. The test also suffers low power under empirically plausible magnitudes of measurement error and likely lengths of follow-up, even when the constant-rate assumption is true. Thus, whilst a significant result of the proposed test is likely to indicate that true lengthening is present in a data set, a non-significant result does not mean that true lengthening is absent.

Keywords: telomere length; biomarkers; statistics; telomere lengthening; aging

Version of December 2016

[1] The authors contributed equally to this paper. Correspondence should be addressed to: daniel.nettle@ncl.ac.uk. The research was supported by the National Centre for the Replacement Refinement and Reduction of Animals in Research (NC3Rs) under grant number NC/K000801/1 and the European Research Council (ERC) under grant number AdG 666669 (COMSTAR).

Introduction

Telomere shortening in tissues such as blood has emerged as an important biomarker of ageing (Müezzinler et al. 2013), predictor of future morbidity and mortality (Heidinger et al. 2012; Boonekamp et al. 2013; Rode et al. 2015), and indicator of accumulated adversity (Hau et al. 2015; Bateson 2016). Telomeres are repetitive DNA sequences at the end of eukaryotic chromosomes that shorten with age. In longitudinal studies, although measured telomere length becomes shorter over successive time points on average, there is often a substantial fraction of the sample that shows an increase in measured telomere length (Steenstrup et al. 2013a; Simons et al. 2014). The observation of apparent lengthening is potentially important, since it points to the possibility that a component of cellular ageing might under some circumstances be reversible in vivo. However, telomere length cannot be measured with perfect precision. There is error variation both due to sampling (heterogeneity in cells within an individual lead to variable estimates of that individual's average telomere length), and measurement (laboratory assays do not produce identical results each time even with the same sample). The existence of error variation means that the second of two longitudinal samples may show a higher value than the first even if the true average telomere length has not increased. Thus, it is possible that apparent telomere lengthening in a sample represents no more than error (Steenstrup et al. 2013a; Bateson and Nettle 2016).

Simons, Stulp and Nakagawa (Simons et al. 2014) recently proposed a statistical test for detecting when there is more observed lengthening in a longitudinal sample than should be expected under the hypothesis of error alone, and hence for inferring when true lengthening is likely to be present in some subset of the sample. This is potentially a useful innovation as it might allow resolution of whether apparent telomere lengthening over time in vivo is a biologically real phenomenon or not. The test requires that each individual is measured at three or more time points. To complete the test, a ratio of two variance estimators (henceforth, the F-ratio) is compared to an F-distribution, in a similar manner to the F-test familiar from ANOVA. Under the null hypothesis (no true lengthening), the two estimators will be similar, the F-ratio will be close to 1, and the *p*-value from comparing the statistic to the F-distribution with appropriate degrees of freedom will be large (i.e. not significant). Under the alternative hypothesis (true lengthening is present), the numerator will be substantially larger than the denominator, the F-ratio will be larger than 1, and the *p*-value will therefore be small (considered significant by the usual convention when $p < 0.05$).

The numerator of the F-ratio estimates the variability in the sample by a calculation based on the number of individuals who have a higher measured telomere length at the final time point compared to the first, and the magnitude of their apparent increase (SSN, equation 5; see SSN, Appendix for derivation of this estimator). The denominator of the F-ratio estimates what under the null hypothesis is the same variability, in a different way. It fits a separate regression line through the points corresponding to the repeat measurements of each individual (so the number of regression lines is equal to the number of individuals in the sample). For each of these lines, it calculates the variance of the residuals, the deviations of the points from the fitted line. This is why three measurement points are required: with just two points, the line goes through both and there is no residual. Finally, the variability of the whole sample is estimated as the mean of the residual variance from each of the separate individual regressions (see SSN, equations 1-3).

There is an important assumption involved in the specification of the denominator of the F-ratio statistic, namely that each individual's telomeres truly change at a constant rate over time. Thus, any deviation of the individual's successive measurement points from a straight line (either going up, going down, or flat) can be taken to represent sampling or measurement error. However, this assumption is biologically unrealistic. The pace of telomere shortening has been linked to infection (Asghar et al. 2015), adverse life events and stress (Epel et al. 2004; Puterman et al. 2014), and health behaviours (Puterman et al. 2014). All of these factors are episodic or changeable over time, so it is plausible that an individual's telomeres change at different rates—or even in different

directions—in different years, without this being in any sense due to measurement or sampling error. If there are year-to-year changes in individuals' rate of true shortening, then the linear regressions for each individual would not fit perfectly even if telomere length could be measured with no error at all. The denominator of the F-ratio statistic proposed by SSN thus actually sums together two components: the variability over time of the *true* rate of telomere change within individuals, plus the measurement and sampling error. This means that, where there is any variability in individual shortening rates over time, the denominator of the test will be larger than it should be for the purposes required of it, the F-ratio will consequently be too small, and the test will potentially be non-significant even when true lengthening is present.

It is common for statistical tests to rely in their derivation on assumptions that are not exactly met in real phenomena, but yet turn out to be useful. Thus, the question is to what extent departures from constant rates of shortening within individuals in realistic datasets cause the proposed test to fail. To investigate this problem, we used a recently-developed computational model of telomere dynamics (Bateson and Nettle 2016) to generate simulated longitudinal data. This model is parameterized with values derived from the empirical literature on human telomeres. Importantly, the model first simulates the true dynamics, then adds empirically-derived levels of measurement error to all of the values. Thus, it is possible using this model to know what the true extent of lengthening in a simulated dataset is, and what the apparent extent of lengthening is once measurement error has been added. The ideal performance criterion for the SSN F-ratio test is that it should be significant when applied to the post-measurement-error dataset whenever the pre-measurement-error dataset contains a substantial number of true lengtheners. This is similar to the approach taken by SSN themselves when they used simulated data to investigate the power of their test. We investigate the performance of the statistical test first where its derivation assumption of a constant rate of change within each individual is met, and then in scenarios where this assumption does not hold.

The model

The model on which the present work is based is described formally in the Appendix. It is explored more fully, and the numerical values chosen for each parameter justified, in Bateson and Nettle (2016). The R code to generate all the results that follow is available as Supporting Online Material. The model assumes that telomere length is measured every year, and it can be iterated to give as many years of data as required.

In the first stage of the model, the true telomere lengths at each time point for n = 10000 individuals are generated. The baseline telomere lengths are drawn from a normal distribution with mean 7000 base pairs (bp) and standard deviation 700 bp. The second year's telomere lengths are generated by subtracting a normally distributed random amount with mean 30 bp and standard deviation 50 bp. This means that although the average telomere length shortens from baseline to the second year, some individuals truly lengthen. For example, an individual whose attrition is one standard deviation from the mean in the positive direction actually experiences lengthening of 20 bp. The values for the means and standard deviations of baseline telomere length and attrition are drawn from the empirical literature.

In each subsequent year, attrition is repeated, again with a mean of 30 bp and standard deviation of 50 bp. Attrition in each successive year can be made to be correlated with attrition in the previous year (each new year's attrition values are generated from the last using equation 1 of Bateson and Nettle (2016)). The level of autocorrelation is controlled by a parameter r. In the case where $r = 1$, the amount of telomere change, whether shortening or elongation, is constant from year to year. Thus, the r = 1 case captures the assumption made by SSN in the derivation of their statistic. Where r = 0, attrition is completely independent from year to year; an individual with relative fast attrition in one year is just as likely as any other to have slow attrition the next year. Here, we investigate three values of r: r = 1, where the assumptions of SSN's proposed test are met; r = 0, where there is no

individual consistency at all in the rate of telomere change; and $r = 0.5$, where there is partial but not complete individual consistency in the rate of change over time.

In a second stage of the model, measurement error is introduced by assuming that measured telomere length at each time point is an independently generated random sample from a normal distribution with the mean equal to the true telomere length. For the standard deviation of this error distribution, we investigated two values: 2% of the average true telomere length, and 8% of the average true telomere length. These values were chosen to be respectively low and high in the observed range of technical variability in telomere measurements across laboratories, which has been estimated to be 1.4 - 9.5 % (Martin-Ruiz et al. 2014).

We used the model to generate one hundred datasets at each combination of: two to eleven years of follow-up; and autocorrelations of $r = 1$, $r = 0.5$ and $r = 0$. All of these datasets contained true telomere lengthening, though the proportion of true lengtheners varied as functions of both length of follow-up and autocorrelation (Bateson and Nettle 2016). For each dataset, we calculated the F-ratio statistic using the code provided by SSN. We investigated, for each combination of years of follow-up and *r*: first, how many true lengtheners there were in each dataset; and second, how many of the possible 100 F-ratio tests were significant by the conventional criterion of $p < 0.05$.

Results

Figure 1 plots the proportion of times the F-ratio test proposed by SSN produced a significant result, as a function of the number of years of follow-up, and broken down by the autocorrelation of individuals' annual true telomere attritions ($r = 0$, $r = 0.5$ or $r = 1$), and the level of assumed measurement error ($CV = 2\%$ or $CV = 8\%$). There is true telomere lengthening in all the datasets, so the optimal proportion of significant tests would be 1.0 in all cases. The mean proportion of individuals whose telomeres truly lengthen does however vary as a function of *r* and the length of follow-up. It is shown as the lower of the two reference lines in each panel.

For perfect autocorrelation ($r = 1$), just over one quarter of the individuals in each dataset shows true lengthening under our assumptions (because the slope of change is constant for each individual when $r = 1$, those that start off by lengthening continue to do so forever; thus the proportion of lengtheners remains constant as the number of years of data increases). Where $r = 1$ and measurement error is small ($CV = 2\%$), the F-ratio test is often significant. Indeed, as the number of years of data increases, the proportion of times the test is significant gradually increases, though at eleven years of follow-up, the power is still less than 100%. Where $r = 1$ and measurement error is large ($CV = 8\%$), the proportion of significant tests is very low at all lengths of follow-up.

When $r = 0$ and $r = 0.5$, the datasets also contain substantial numbers of true lengtheners. With these values for autocorrelation, the proportion of true lengtheners decreases with increasing length of follow-up. This is because, under these assumptions, individuals with an initially lengthening trajectory tend to regress towards the expected value of telomere change (i.e. gradual attrition) with the passage of more time. Nonetheless, even at the longest follow-up periods, the proportion of true lengtheners in these datasets is substantially greater than zero. For autocorrelation values of $r = 0$ and $r = 0.5$, the F-ratio test does not detect lengthening; it is significant at most 1-2% of the time, regardless of the length of follow-up.

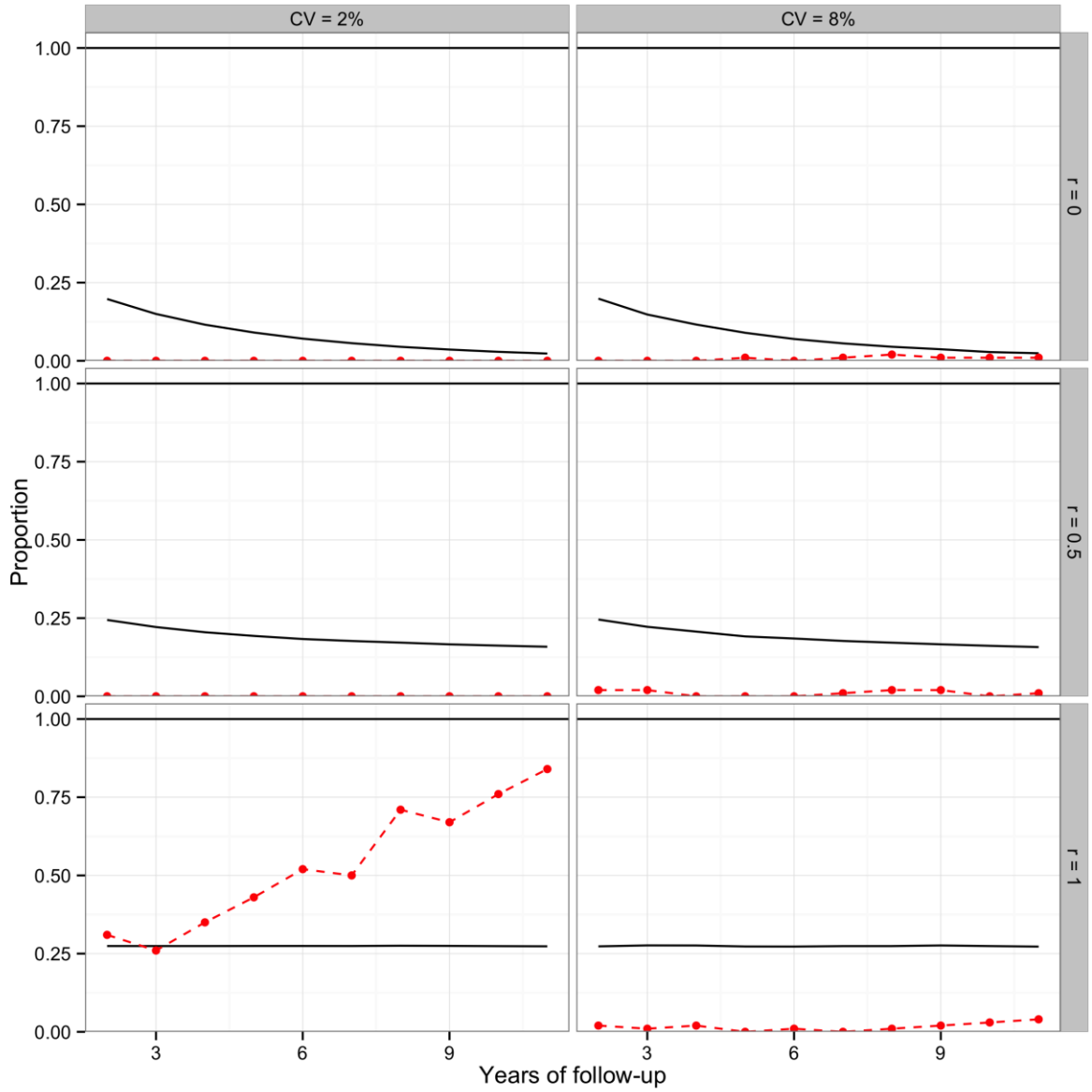

Figure 1. The proportion of times the F-ratio test proposed by Simons, Stulp and Nakagawa (2014) returned a significant result (points and dashed lines) for different numbers of years of follow-up, split by level of measurement error (*CV* =2% or *CV* = 8%), and values of the autocorrelation parameter *r* (*r* = 0, *r* = 0.5, *r* = 1). Given that some true elongation was present in all datasets, the proportion of times the test was significant is an estimate of its statistical power. When *r* = 1, individuals have a constant rate of change over the whole time period. When *r* = 0, an individual's telomere change in one time period is independent of their change in the previous period. At each combination of *r,* CV, and years of follow-up, 100 datasets each of 10,000 individuals were simulated. On each panel, the lower solid line represents the mean proportion of individuals that were true lengtheners at that combination of parameters.

Discussion

We considered the performance of the F-ratio test proposed by SSN on longitudinal datasets simulated using empirically-derived parameters, where there was a non-zero and known proportion of true telomere lengtheners, and the sample size was very large. Thus, ideally the test should have been significant in all or the vast majority of cases.

Where the autocorrelation parameter *r* was set to 1, individuals had a constant rate of change from one year to the next, and hence the derivation assumptions of the test were met. Here, the test was indeed often significant as long as measurement error was small relative to the magnitude of true attrition. The test's power to detect true lengthening was very low, however, when measurement error was larger than the variation in true attrition. This result is in accordance with those presented by SSN in their analysis of the test's power. They found that power was good as long as the standard deviation of true attrition was larger than the standard deviation due to measurement error (see SSN, figure 1). However, the empirical best estimates are that the standard deviation of annual true telomere attrition is of the order 14 - 53 bp/year (Aviv et al. 2009; Chen et al. 2011; Kark et al. 2012; Steenstrup et al. 2013b), whilst the standard deviation due to measurement error is of the order of 98 - 665 bp (Martin-Ruiz et al. 2014; Bateson and Nettle 2016). Although this can be reduced by running additional technical replicates, the reality is that measurement error is likely to be at least as large as the variation in true telomere change in many empirical situations.

We found the power of the test to be modest when the follow-up time was shorter than around 8 years, even assuming perfect autocorrelation and the low value of measurement error (and bearing in mind that the sample size was 10,000 individuals per dataset in our simulations). This appears to contrast with the findings by SSN that the power of the test is high in simulated datasets as small as 400 individuals sampled at three time points. However, SSN do not specify what the intervals between their time points represent. They assume an average loss of telomere length from one time point to the next of 3% of the starting length. Given that human telomere length is around 7000 bp and estimated annual attrition is around 30 bp (Aviv et al. 2009; Müezzinler et al. 2013), each of SSN's time intervals can be equated to around 7 years in adult humans, meaning that their scenario is equivalent to 14 years of follow-up in total. Thus, their results and ours are in accordance that the power of the test is good when its assumption of constant individual attrition rate is met, measurement error is small relative to variability in true attrition and the length of follow-up is of the order of a decade or more.

When individuals' true attrition rates were allowed to vary from year to year, the test failed catastrophically. In all of our scenarios with $r = 0$ or $r = 0.5$, some fraction of the individuals in each dataset—sometimes up to 25% -- were true lengtheners, and yet the test returned a significant result less than the one time out of twenty that would be expected by chance alone under the null hypothesis. This was true at all lengths of follow-up. Because the denominator of the F-ratio, by adding the variability in the annual rate of attrition to the calculation of the error variation, is systematically too large, the F-ratio statistic is almost always less than one (whether the null hypothesis is true or false), and a significant result can very rarely be generated. Thus, it seems that the assumption of a constant rate of change for each individual is unfortunately an instance of a derivational assumption whose violation in the empirical world does turn out to matter for the usability of the statistic in practice.

We do not know the correct value for the autocorrelation of true individual telomere length change over time in human populations. Bateson and Nettle (2016) used observed patterns of apparent lengthening in data sets with different durations of follow-up to estimate that it is low. Several recent empirical studies have suggested that telomere change tends to oscillate, with periods of rapid attrition followed by periods of elongation (Svenson et al. 2011; Huzen et al. 2014). This would suggest low or even negative autocorrelation values. However, in these studies, the true dynamics

are not clearly distinguished from the effects of measurement error, which would be capable of producing apparent oscillations due to regression to the mean.

Nonetheless, although we are uncertain about what value of the temporal autocorrelation in individual telomere change is correct, it is an extremely restrictive and unjustified to assume it to be perfect. Indeed, what attracts researchers to telomere length as a biomarker is precisely that the rate of attrition seems to vary in relation to life events (Epel et al. 2004; Shalev 2012; Asghar et al. 2015; Bateson 2016). Thus, the interpretation of the result of the SSN test for true lengthening should be extremely cautious. Although a significant test result probably does indicate the presence of true lengthening, a non-significant test result does not indicate the absence of true lengthening. Instead, it may simply indicate that the autocorrelation of true individual change in the sample is not perfect; that measurement error is larger relative to the true variability in the rate of telomere attrition; or else that the autocorrelation is perfect and the measurement error sufficiently small, but the follow-up period is too short relative to the samples size and measurement for the test to yield a significant result. These uncertainties mean that the test is unlikely to have any practical application.

The objective of the SSN test is an extremely useful one, since it would be of interest to be able to simply resolve whether apparent telomere lengthening in longitudinal samples always represents error, and also to identify those individuals whose telomeres have lengthened for further study. Given the uncertainty about the true variability of individuals' rates of telomere change, we cannot propose any simple correction to the test, or alternative test statistic. Instead, we suggest that investigators who have access to high-quality longitudinal datasets on human telomere change use those datasets to estimate and report key parameters germane to our model, such as the mean and standard deviation of annual attrition, the coefficient of error variation (which can be estimated from running technical replicates of the same samples, or better, multiple samples from the same individual taken on the same day), and the autocorrelation of individual telomere change from repeated samples (which will have to be corrected for measurement error). With these values in hand, it will be possible to create a fairly accurate process model of the underlying true dynamics for each sample. Amongst other things, this will allow estimation of what proportion, if any, of individuals shows true telomere lengthening.

Appendix

For each individual in each dataset, baseline telomere length in base pairs is generated by:

$$length_b \sim N(7000, 700) \quad (1)$$

Length at the first follow-up year is then generated by:

$$length_1 = length_b - attrition_1 \quad (2)$$

$$attrition_1 \sim N(30, 50) \quad (3)$$

For all subsequent years:

$$length_{y+1} = length_y - attrition_{y+1} \quad (4)$$

$$attrition_{y+1} = r \cdot attrition_y + \sqrt{(1-r^2)}\, N(\frac{(1-r)}{\sqrt{(1-r^2)}}30, 50) \quad (5)$$

Equation 5 generates attrition values that have the required level of autocorrelation *r*, whilst maintaining a mean attrition of 30 bp and a standard deviation of attrition of 50 bp (for proof see Bateson and Nettle 2016).

Finally, measurement error is added to all telomere lengths using:

$$measured_y \sim N(length_y, length_y * \frac{CV}{100}) \quad (6)$$

Here, *CV* is a coefficient of variation due to measurement error, taken as either 2 or 8, as specified.

References

Asghar M, Hasselquist D, Zehtindjiev P, et al (2015) Hidden costs of infection: Chronic malaria accelerates telomere degradation and senescence in wild birds. 347:9–12.

Aviv A, Chen W, Gardner JP, et al (2009) Leukocyte telomere dynamics: Longitudinal findings among young adults in the Bogalusa Heart Study. Am J Epidemiol 169:323–329. doi: 10.1093/aje/kwn338

Bateson M (2016) Cumulative stress in research animals: Telomere attrition as a biomarker in a welfare context? BioEssays 38:201–12. doi: 10.1002/bies.201500127

Bateson M, Nettle D (2016) The telomere lengthening conundrum - it could be biology. Aging Cell 2016:1–8. doi: 10.1111/acel.12555

Boonekamp JJ, Simons MJP, Hemerik L, Verhulst S (2013) Telomere length behaves as biomarker of somatic redundancy rather than biological age. Aging Cell 12:330–2. doi: 10.1111/acel.12050

Chen W, Kimura M, Kim S, et al (2011) Longitudinal versus cross-sectional evaluations of leukocyte telomere length dynamics: Age-dependent telomere shortening is the rule. Journals Gerontol - Ser A Biol Sci Med Sci 66 A:312–319. doi: 10.1093/gerona/glq223

Epel ES, Blackburn EH, Lin J, et al (2004) Accelerated telomere shortening in response to life stress. Proc Natl Acad Sci U S A 101:17312–5. doi: 10.1073/pnas.0407162101

Hau M, Greives TJ, Haussmann MF, et al (2015) Repeated stressors in adulthood increase the rate of biological ageing. Front Zool 12:1–10. doi: 10.1186/s12983-015-0095-z

Heidinger BJ, Blount JD, Boner W, et al (2012) Telomere length in early life predicts lifespan. Proc Natl Acad Sci U S A 109:1743–8. doi: 10.1073/pnas.1113306109

Huzen J, Wong LSM, Veldhuisen DJ Van, et al (2014) Telomere length loss due to smoking and metabolic traits. J Intern Med 275:155–163. doi: 10.1111/joim.12149

Kark J, Goldberger N, Kimura M, et al (2012) Energy intake and leukocyte telomere length in young adults. Am J Clin Nutr 479–87. doi: 10.3945/ajcn.111.024521.1

Martin-Ruiz CM, Baird D, Roger L, et al (2014) Reproducibility of telomere length assessment: an international collaborative study. Int J Epidemiol 1–11. doi: 10.1093/ije/dyu191

Müezzinler A, Karina A, Brenner H (2013) A systematic review of leukocyte telomere length and age in adults. Ageing Res Rev 12:509–519. doi: 10.1016/j.arr.2013.01.003

Puterman E, Lin J, Krauss J, et al (2014) Determinants of telomere attrition over 1 year in healthy older women: stress and health behaviors matter. Mol Psychiatry 1–7. doi: 10.1038/mp.2014.70

Rode L, Nordestgaard BG, Bojesen SE (2015) Peripheral blood leukocyte telomere length and mortality among 64,637 individuals from the general population. J Natl Cancer Inst 107:djv074. doi: 10.1093/jnci/djv074

Shalev I (2012) Early life stress and telomere length: Investigating the connection and possible mechanisms: A critical survey of the evidence base, research methodology and basic biology. BioEssays 34:943–52. doi: 10.1002/bies.201200084

Simons MJP, Stulp G, Nakagawa S (2014) A statistical approach to distinguish telomere elongation from error in longitudinal datasets. Biogerontology 15:99–103. doi: 10.1007/s10522-013-9471-2

Steenstrup T, Hjelmborg JVB, Kark JD, et al (2013a) The telomere lengthening conundrum--artifact or biology? Nucleic Acids Res 41:e131. doi: 10.1093/nar/gkt370

Steenstrup T, Hjelmborg JVB, Mortensen LH, et al (2013b) Leukocyte telomere dynamics in the elderly. Eur J Epidemiol 28:181–187. doi: 10.1007/s10654-013-9780-4

Svenson U, Nordfjall K, Baird D, et al (2011) Blood cell telomere length is a dynamic feature. PLoS One 6:e21485. doi: 10.1371/Citation