test performed here and its *P*-value. Furthermore, in the Discussion it is stated that 'inter-laboratory variation is no longer different between the techniques if calculated on the basis of z-scored data.' Our calculations do not confirm this statement.

We do not believe that any of the errors we found were intentional. We believe that the paper demonstrates the need for more double-checking and better statistical analysis.

We think that the four Key Messages of the paper still hold. However, in our opinion the first key message, 'Rankings are very similar if different laboratories measure telomere lengths in the same samples', is now too strong and instead of 'very similar' we would suggest changing it to just 'similar'. Furthermore, the authors of the paper in the conclusion say that: 'Z-scoring of data appears at present the best possibility for combining results from different laboratories' and it is based on wrongly calculated z-scores.

We think that the difference of measured telomere lengths between different laboratories and techniques is an important issue. We think that the authors tackled the issue with the right design of experiment and that they produced findings valuable to the scientific community. It is unfortunate that their statistical analysis had so many errors and thus questions some of their findings.

Supplementary Data

Supplementary data are available at *IJE* online.

References

1. Martin-Ruiz CM, Baird D, Roger L *et al*. Reproducibility of telomere length assessment: an international collaborative study. *Int J Epidemiol* 2015;**44:**1673–83.
2. Friedman M. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc.* 1937;**32**:675–701.
3. Wilcoxon F. Individual comparisons by ranking methods. *Biometrics Bull* 1945;**1**:80–83.

Reproducibility of telomere length assessment: Authors' Response to Damjan Krstajic and Ljubomir Buturovic

International Journal of Epidemiology, 2015, 1739–1741
doi: 10.1093/ije/dyv170
Advance Access Publication Date: 24 September 2015

From Carmen M Martin-Ruiz,[1] Duncan Baird,[2] Laureline Roger,[2] Petra Boukamp,[3] Damir Krunic,[3] Richard Cawthon,[4] Martin M Dokter,[5] Pim Van Der Harst,[5] Sofie Bekaert,[6] Tim De Meyer,[13] Goran Roos,[7] Ulrika Svenson,[7] Veryan Codd,[8] Nilesh J Samani,[8] Liane Mcglynn,[9] Paul G Shiels,[9] Karen A Pooley,[10] Alison M Dunning,[11] Rachel Cooper,[12] Andrew Wong,[12] Andrew Kingston[1] and Thomas Von Zglinicki[1]*

[1]Institute for Ageing and Health, Newcastle University, Newcastle, UK, [2]Institute of Cancer and Genetics, Cardiff University, Cardiff, UK, [3]Deutsches Krebsforschungszentrum (DKFZ), Heidelberg, Germany, [4]Department. of Human Genetics, University of Utah, Salt Lake City, UT, USA, [5]Department of Cardiology, University of Groningen, Groningen, The Netherlands, [6]Bimetra, Clinical Research Center, Ghent University Hospital, Ghent, Belgium, [7]Department of Medical Biosciences, Umeå University, Umeå, Sweden, [8]Department of Cardiovascular Sciences, University of Leicester, Leicester, UK, [9]Institute of Cancer Sciences, University of Glasgow, Glasgow, UK, [10]Centre for Cancer Genetic Epidemiology, Department of Public Health and Primary Care, University of Cambridge, Cambridge, UK, [11]Centre for Cancer Genetic Epidemiology, Department of Oncology, University of Cambridge, Cambridge, UK, [12]MRC Unit for Lifelong Health and Ageing, University College London, London, UK and [13]Department of Mathematical Modelling, Statistics and Bioinformatics, Ghent University, Ghent, Belgium

*Corresponding author. E-mail: t.vonzglinicki@newcastle.ac.uk

In a recent comment, Krstajic and Buturovic[1] checked the calculations which we did in our paper.[2] We are very grateful to them for spotting two inconsistencies in our handling of the data. We apologize for these and have corrected them in an accompanying Corrigendum.[3] The corrected calculations (see Figure 1 below and Corrigendum[3]) confirmed our previous conclusions, with the sole exception that some rank correlation coefficients between laboratories (specifically involving STELA) were lower than appreciated before. Therefore, we agree with the suggestion by Krstajic and

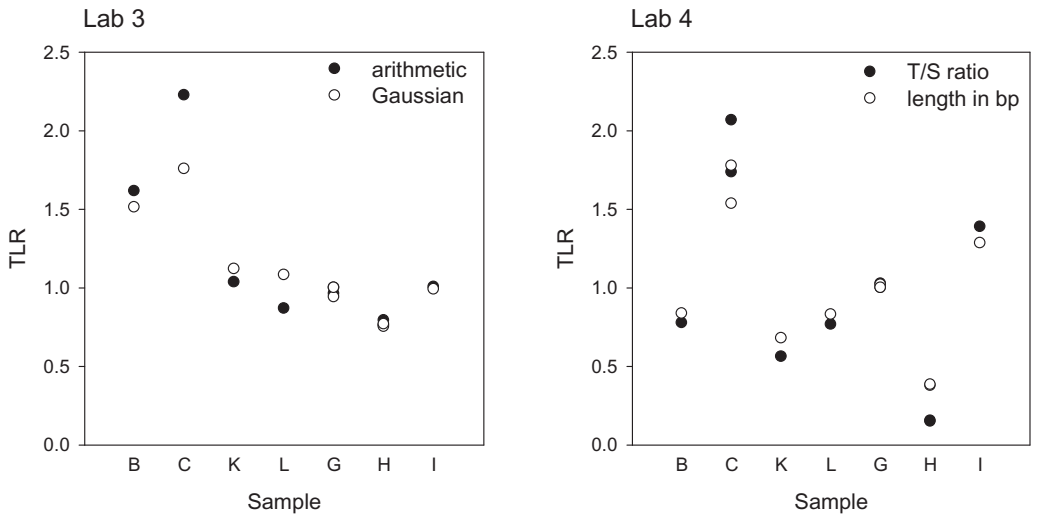

Figure 1. TLRs for Labs 3 and 4 (round 2) calculated from two sets of raw values.

Buturovic[1] that we should tone down our conclusion of 'very similar rankings' between laboratories to 'similar rankings' (see Corrigendum[3]). However, we reject multiple arguments by Krstajic and Buturovic[1] (see below). Specifically, we maintain our conclusion that 'inter-laboratory variation is no longer different between the techniques if calculated on the basis of z-scored data', because their criticism of this statement is based on processing errors in their data.

None of the corrections to the original paper[3] changes in any way the arguments presented in our response[4] to a recent independent comment by Verhulst *et al.*[5]

In detail, we respond to the points raised[1] as follows.

(i) *Telomere length ratios (TLR) values for Laboratories 3 and 4 in round 2, as given in Table 2, are different from those calculated based on the raw values given in Supplementary Table S2.* This is correct. The error occurred because the raw telomere length data returned from Laboratories 3 and 4 were each calculated in two different ways. Lab 3 used either arithmetic means or a Gaussian fit to calculate telomere length from the positions of the bands on the gels, and Lab 4 calculated telomere length either as T/S ratios or as absolute length in bp using an internal normalization. In order not to over-complicate our paper,[2] we chose to present data from only one calculation method per laboratory. Data in Table 2, round 2, were calculated from the arithmetic fit for Lab 3 and from T/S ratios for Lab 4. Unfortunately, raw data given in Supplementary Table S2 round 2 in our paper[2]were the alternative data sets for Labs 3 and 4, i.e. results from Lab 3 calculated by Gaussian fit and from Lab 4 calculated as absolute length using internal normalization. In the Corrigendum[3] we have now re-performed all calculations based on the raw data given in Supplementary Table S2 in the original paper.[2] The differences in TLRs as calculated from both data sets are generally small, as shown in Figure 1, especially if compared with the inter-lab differences as shown in Figure 1 of our original paper.[2] With the exception of the strength of rank correlations between different laboratories (see point 2 below), all conclusions remain unchanged.

(ii) *In Supplementary Table S3, wrong correlation coefficients are given.* This is correct and we apologize. Pearson correlation coefficients instead of Spearman's rank correlation coefficients were mistakenly given in the table. Correct Spearman's rank correlation coefficients are given in the Corrigendum.[3] The range of correlation coefficients is now 0.25–0.99, due to STELA results correlating less well with both Southern and qPCR results after Gaussian data fitting. Therefore, we agree with Krstajic and Buturovic[1] and change our first key message to 'Rankings are similar if different laboratories measure telomere lengths in the same samples' (see Corrigendum[3]).

(iii) *Comparison of intra-batch coefficients of variation (CVs).* We do not understand how Krstajic and Buturovic[1] arrived at a *P*-value of 0.784 for an analysis of variance (ANOVA) comparing intra-batch CVs between laboratories. We checked the value given in our paper[2] ($P = 0.299$). It remains unchanged when calculations are based on the set of raw data for Labs 3 and 4, as given in Supplementary Table S2. Depending on how the two datasets for Lab 10 are included in the comparison laboratory, a Kruskal-Wallis test becomes appropriate, but this again will not yield a $P = 0.784$. None of the outcomes is significant.

(iv) *Should inter-laboratory comparisons based on intra- and inter-batch CVs be evaluated using tests for dependent (paired) data?* Treating the data as paired would have been

correct if we, as Krstajic and Buturovic[1] seem to assume, were comparing directly telomere lengths. We would then separately assess the capability of individual laboratories to measure, say, long vs short telomeres. However, our study was neither intended nor powered sufficiently to do this. We were not interested whether a certain laboratory might be better in measuring a certain type of telomeres, whether long or short, tumour or lymphocytes etc. What we were interested in was the capability of the laboratories and, especially of the different techniques, to measure an essentially random set of telomeres reproducibly. In that respect it was not relevant whether the same or different DNA samples were used. In fact, 'real' samples would not be paired between laboratories. A paired analysis design would therefore over-interpret differences between laboratories or techniques. Therefore we were using CVs, which normalized the results against the most relevant difference between samples (i.e. their mean telomere length) and treated the individual samples as random (which is how they were measured in our fully blinded design).

(v) *Why are z-scores and their inter-lab variations different between Krstajic and Buturovic[1] and our paper[2]?* Krstajic and Buturovic[1] calculated z-scores independently for rounds 1 and 2. We calculated z-scores per laboratory, using a common average between both rounds. We think our approach is the more appropriate because it has two advantages: it focuses on the performance of the laboratories (which we wanted to compare) irrespective of the differences between rounds 1 and 2, and it gives better statistical power for the calculation of the scores. In any case, these two approaches result only in minor differences (compare data for Labs 1 to 9 in Supplementary Table ST3 from Krstajic and Buturovic[1] with our Supplementary Table S4[2]).

Importantly, Krstajic and Buturovic[1] claim that their calculations do not confirm our statement that 'inter-laboratory variation is no longer different between the techniques if calculated on the basis of z-scored data'. This claim, however, is based on a calculation error in their Supplementary Table 3[1]: they did not convert TLRs for Labs 10-1 and 10-2 into z-scores, resulting in artificially large differences in the inter-lab variation.

(vi) *Key messages and conclusions*. In agreement with-Krstajic and Buturovic[1] we change our first key message to: 'Rankings are similar if different laboratories measure telomere lengths in the same samples'. We retain our conclusion that: 'Z-scoring of data appears at present the best possibility for combining results from different laboratories', because it is based on correctly calculated z-scores.

References

1. Krstajic D, Buturovic L. Reproducibility of telomere length assessment. *Int J Epidemiol* 2015;**44:**1738–39.
2. Martin-Ruiz CM, Baird D, Roger L *et al*. Reproducibility of telomere length assessment: an international collaborative study. *Int J Epidemiol* 2015;**44:**1673–83.
3. Martin-Ruiz CM, Baird D, Roger L *et al*. Reproducibility of telomere length assessment: an international collaborative study. Corrigendum. *Int J Epidemiol* 2015;**44:**1749–54.
4. Martin-Ruiz CM, Baird D, Roger L *et al*. Is Southern blotting necessary to measure telomere length reproducibly? *Int J Epidemiol* 2015;**44:**1686–87.
5. Verhulst S, Susser E, Faktor-Litvak PR *et al*. The reliability of telomere length measurements. *Int J Epidemiol* 2015;**44:** 1683–86.

Childhood cancer—the role of birthweight and antenatal radiography

International Journal of Epidemiology, 2015, 1741–1743
doi: 10.1093/ije/dyv158
Advance Access Publication Date: 11 August 2015

From Richard Wakeford[1]* and John F Bithell[2]

[1]Institute of Population Health, The University of Manchester, Manchester, UK and [2]Department of Statistics, University of Oxford, Oxford, UK

*Corresponding author. Centre for Occupational and Environmental Health, Institute of Population Health, The University of Manchester, Ellen Wilkinson Building, Oxford Road, Manchester M13 9PL, UK. E-mail: richard.wakeford@manchester.ac.uk

The recent study of O'Neill and her colleagues[1] increases the evidence[2–6] for a raised risk of childhood leukaemia (and some other childhood cancers) with increased birthweight. With funding from Children with Cancer UK, we