



*Citation for published version:*

Dumuid, D, Stanford, TE, Martin-Fernández, JA, Pedisic, Z, Maher, C, Lewis, L, Hron, K, Katzmarzyk, PT, Chaput, J-P, Fogelholm, M, Hu, G, Lambert, EV, Maia, J, Sarmiento, OL, Standage, M, Barreira, TV, Broyles, ST, Tudor-Locke, C, Tremblay, MS & Olds, T 2018, 'Compositional data analysis for physical activity, sedentary time and sleep research', *Statistical Methods in Medical Research*, vol. 27, no. 12, pp. 3726-3738.  
<https://doi.org/10.1177/0962280217710835>

*DOI:*

[10.1177/0962280217710835](https://doi.org/10.1177/0962280217710835)

*Publication date:*

2018

*Document Version*

Peer reviewed version

[Link to publication](#)

Dorothea Dumuid, Tyman E Stanford, Josep-Antoni Martin-Fernández, Željko Pediši, Carol A Maher, Lucy K Lewis, Karel Hron, Peter T Katzmarzyk, Jean-Philippe Chaput, Mikael Fogelholm, Gang Hu, Estelle V Lambert, José Maia, Olga L Sarmiento, Martyn Standage, Tiago V Barreira, Stephanie T Broyles, Catrine Tudor-Locke, Mark S Tremblay, Timothy Olds, Compositional data analysis for physical activity, sedentary time and sleep research, *Statistical Methods in Medical Research*. Copyright © 2017 The Author(s). Reprinted by permission of SAGE Publications.

## University of Bath

### General rights

Copyright and moral rights for the publications made accessible in the public portal are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

### Take down policy

If you believe that this document breaches copyright please contact us providing details, and we will remove access to the work immediately and investigate your claim.

## 1 **1 Introduction: Daily activity behaviours are compositional by nature**

2

3 Physical inactivity is considered to be a major risk factor for non-communicable disease  
4 and premature death.<sup>1</sup> A global economic analysis has estimated the health-care system  
5 cost of physical inactivity in 2013 alone to be \$ (INT\$) 53.8 billion.<sup>2</sup> To produce such  
6 economic estimates, physical inactivity is defined as relatively lower amounts of  
7 moderate-to-vigorous-intensity physical activity (MVPA), but underlying analyses fail  
8 to account for the fact that when time in MVPA is reduced, a subsequent and equal  
9 increase in time must be distributed to the remaining behaviour domains: sleep,  
10 sedentary time and light-intensity physical activity (light PA), to represent the finite 24-  
11 hours, or 1440 minutes, in any given day. These other non-MVPA behaviours may  
12 themselves have positive or negative effects on health and mortality.<sup>3-6</sup> Therefore, the  
13 health and economic burden of physical inactivity per se remains unclear. Similar  
14 estimates for sedentary time<sup>7</sup> are uncertain for the same reason, i.e., they fail to  
15 adequately account for other behaviours. Pedišić<sup>8</sup> argued that studies on health  
16 outcomes of sleep duration and light PA can be put under the same scrutiny.

17

18 Time spent in MVPA represents one of the exhaustive and mutually exclusive  
19 components of an individual's 24-h day. The non-MVPA time remaining within an  
20 individual's day can be partitioned into light PA, sedentary time and sleep and all of

1 them can be considered as relative contributions to the overall time budget. The defined  
2 behaviours (MVPA, light PA, sedentary time and sleep) are therefore compositional  
3 data, and have important properties that must be respected.<sup>9</sup>  
4  
5 Consider a vector  $\boldsymbol{x} = [x_1, x_2, \dots, x_D] \in \mathbb{R}^D$  with positive components, where  $\sum x_i = C$ ,  
6 and  $C$  is the closure constant. The sample space of the vector  $\boldsymbol{x}$  can thus be represented  
7 by the D-part simplex ( $S^D$ ), which is a ( $D-1$ )-dimensional subset of the real space ( $\mathbb{R}^D$ )  
8 due to the constant sum constraint of  $C$ . Compositional data are scale invariant<sup>9</sup> because  
9 the application of a common factor  $a$  to the parts  $x_i$  where  $i = 1, 2, \dots, D$  ensures the  
10 relative difference between the parts is maintained, as  $\sum_{i=1}^D ax_i = a \sum_{i=1}^D x_i = aC$ . The  
11 numerical value of the closure constant (e.g., 24 h, one week, one month) is irrelevant.  
12 Daily behaviours could equivalently be measured in hours, minutes or percentages as  
13 the data convey only relative information. The property of scale invariance means  
14 compositional data are in fact elements of equivalence classes of proportional vectors.<sup>10</sup>  
15 Accordingly, the simplex is the sample space of representatives of compositional data  
16 with a chosen constant sum constraint. Specific properties of compositional data are  
17 followed by a natural geometry, known as the *Aitchison geometry*.<sup>11</sup> The closure  
18 constant representation of compositions imposes perfect multi-collinearity among the  
19 components, causing the covariance structure of the data to be negatively biased.<sup>9</sup>  
20 Accordingly, traditional statistical methods for unconstrained variables in real space

1 (e.g., t-tests, multiple linear regression) are not predicative with respect to the specific  
2 geometry of the simplex sample space, and should not be used for absolute or raw  
3 measures of time spent in daily behaviours.<sup>12</sup>

4

## 5 **2 The log-ratio approach for compositional data analysis**

6 The invalidity of standard multivariate techniques for analyzing untransformed or raw  
7 compositional data was recognized in scientific fields decades ago,<sup>13</sup> and in the 1980s  
8 Aitchison proposed a new methodology for the analysis of compositional data.<sup>9</sup> The  
9 methodology is based on the premise that any composition (e.g., an individual's daily  
10 time budget) can be expressed in terms of ratios of its parts (e.g., duration of sleep,  
11 sedentary time, light PA and MVPA). The expression of compositional data as log-ratio  
12 coordinates transfers them from the constrained simplex space to the unconstrained real  
13 space, where traditional multivariate statistics may be applied.<sup>14</sup> The presence of zeros  
14 in a compositional dataset prohibits applying log-ratio coordinates. Several methods  
15 have been proposed to deal with zeros;<sup>15</sup> however they are beyond the scope of this  
16 paper.

17

18 A number of log-ratio coordinate systems for compositional data have been described.<sup>11</sup>

19 One such coordinate system, the additive log-ratio (*alr*), has coordinates,  $\mathbf{a}$ , defined by

20 
$$\mathbf{a} = [a_1, \dots, a_{D-1}] = alr(\mathbf{x}) = [\ln\left(\frac{x_1}{x_D}\right), \ln\left(\frac{x_2}{x_D}\right), \dots, \ln\left(\frac{x_{D-1}}{x_D}\right)]. \quad (1)$$

1 However, the *alr* coordinates are asymmetric, because the components  $x_1, x_2, \dots, x_{D-1}$  are  
 2 divided by the component  $x_D$ . Moreover, they are not isometric, i.e., distances and  
 3 angles in the Aitchison geometry are violated by using the *alr* coordinates, limiting their  
 4 use in statistical applications. This means that the system of *alr* coordinates in the  
 5 Aitchison geometry is oblique, and traditional statistical methods which assume  
 6 orthogonal coordinates therefore cannot be directly applied. Another coordinate system  
 7 is the centred log-ratio (*clr*) coordinate system,  $\mathbf{c}$ , defined as

$$8 \quad \mathbf{c} = [c_1, \dots, c_D] = \text{clr}(\mathbf{x}) = \left[ \ln\left(\frac{x_1}{\tilde{x}}\right), \ln\left(\frac{x_2}{\tilde{x}}\right), \dots, \ln\left(\frac{x_D}{\tilde{x}}\right) \right], \quad (2)$$

9 where  $\tilde{x}$  is the geometric mean of all the D components of the vector  $\mathbf{x}$ . The *clr* are  
 10 symmetric and isometric; they produce a singular covariance matrix because  $\sum_{j=1}^D c_j =$   
 11 0. The *clr* are, strictly speaking, not coordinates but coefficients with respect to a  
 12 generating system. The covariance matrix of *clr* coefficients is singular, so the *clr*  
 13 coefficients cannot be fully utilized as independent variables in multiple regression  
 14 analysis.

15  
 16 The singularity problem of the *clr* can be overcome by the use of an isometric log-ratio  
 17 (*ilr*) coordinate system. Isometric log ratio coordinates form an isometric mapping of  
 18 the composition from the simplex sample space to the real space.<sup>16</sup> To construct the *ilr*  
 19 coordinates, an orthonormal basis coherent with the Aitchison geometry is created in the  
 20  $(D-1)$ -dimensional hyperplane of the *clr* coordinates. Many possible orthonormal

1 coordinate systems can be created, however, following Pawlowsky-Glahn et al.,<sup>11</sup> one  
 2 can define a particular *ilr* system of coordinates through a partitioning process.<sup>16</sup> For  
 3 our purposes, specific *ilr* coordinates based on a sequential partition of one part to the  
 4 remaining compositional parts<sup>17</sup> is very useful. Such *ilr* coordinates result in a  $(D-1)$ -  
 5 dimensional real vector,  $\mathbf{z}$ , defined as

$$\begin{aligned}
 6 \quad \mathbf{z} = [z_1, z_2, \dots, z_{D-1}] = ilr(\mathbf{x}) = & \left[ \sqrt{\frac{D-1}{D}} \ln \left( \frac{x_1}{\sqrt{\prod_{k=2}^D x_k}} \right), \sqrt{\frac{D-2}{D-1}} \ln \left( \frac{x_2}{\sqrt{\prod_{k=3}^D x_k}} \right), \dots \right. \\
 7 \quad & \left. \dots, \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{x_j}{\sqrt{\prod_{k=j+1}^D x_k}} \right), \dots, \frac{1}{\sqrt{2}} \ln \left( \frac{x_{D-1}}{x_D} \right) \right]. \quad (3)
 \end{aligned}$$

8 It can be shown that

$$9 \quad c_1 = \sqrt{\frac{D-1}{D}} z_1, \quad (4)$$

10 i.e., the first *ilr* coordinate in this system is directly proportional to the first *clr*  
 11 coefficient. Both  $c_1$  and  $z_1$  can be interpreted in the same way in terms of dominance of  
 12 a component (here the first compositional part) to the rest of parts. The remaining *ilr*  
 13 coordinates ( $z_2, z_3, \dots, z_{D-1}$ ) contain no relative information regarding the first  
 14 compositional part. If a different compositional part is of interest, e.g., the second part,  
 15 it is simply a matter of rearranging the compositional parts so that the part of interest is  
 16 in the first place, and then reconstructing the *ilr* coordinates according to (3).<sup>18</sup>

17

1 The *ilr* coordinates can also be constructed using a sequential binary partition (SBP) as  
2 described by Egozcue and Pawlowsky-Glahn.<sup>16</sup> The first step in the SBP process  
3 requires division of the full composition into two subgroups of parts, where one  
4 subgroup will form the numerator and the other the denominator of the first *ilr*  
5 coordinate. In the subsequent steps, each of these two subgroups is further split into new  
6 subgroups to create the remaining *ilr* coordinates. For example, based on prior  
7 knowledge of the nature of the behaviour components, the numerator could be selected  
8 to consist of inactivity-related behaviours (sleep and sedentary time) and the  
9 denominator could be activity-related components (light PA and MVPA). The second  
10 SBP then divides the numerator of the first *ilr* coordinate to form the second *ilr*  
11 coordinate, with sleep as the numerator and sedentary time as the denominator. The  
12 final SBP divides the denominator of the first *ilr* coordinate, with light PA as the  
13 numerator and MVPA the denominator of the third *ilr* coordinate.

14

15 As the *ilr* coordinates defined in (3) are orthogonal, the columns of the design matrix  
16 are linearly independent. This avoids the previous issue of a singular covariance matrix  
17 in the multiple linear regression fit. The regression parameters estimated for the first *ilr*  
18 coordinate in the model represent the effect on an outcome when the first component  
19 (numerator) is changed in relation the geometric mean of the remaining parts  
20 (denominator). To examine the influence of the other compositional parts (relative to

1 the geometric mean of the respective remaining parts), a total of  $D$  models are fitted,  
2 with each model including a set of *ilr* which iteratively has a different compositional  
3 part as the numerator of the first *ilr* coordinate (and the remaining parts as the  
4 denominator). The constant term, and other external covariate terms, as well as the  
5 quality of fit, are invariant to the choice of *ilr* basis.<sup>11</sup>

### 7 **3. Compositional data analysis in the field of sleep, sedentary and physical activity** 8 **research**

9 The log-ratio approach for compositional data analysis is well established in many  
10 scientific fields (e.g., geology, biology, hydrology, ecology and economics), and is  
11 considered the gold-standard for analyzing compositional data.<sup>11</sup> However, this  
12 methodology has only recently been used in health research, with researchers applying  
13 compositional data analysis to nutrition,<sup>19</sup> epidemiology<sup>20</sup> and microbiome data.<sup>21</sup>  
14 Furthermore, the compositional nature of daily activity data (sleep, sedentary time, light  
15 PA and MVPA) was not acknowledged until 2014, when Pedišić<sup>8</sup> warned that  
16 traditional analyses within the field were undermined due to the use of inappropriate  
17 and invalid statistical procedures, and called for a paradigm shift towards a  
18 compositional approach. Subsequently, Chastin et al.<sup>18</sup> and Carson et al.<sup>22</sup>  
19 demonstrated the feasibility of estimating the relationship between the complete daily  
20 behaviour composition and health outcomes using the *ilr* methodology outlined above.



1 However, the interpretation of log-ratio regression coefficients is not straight-forward.<sup>23</sup>  
 2 As daily activity data have a meaningful total, i.e., 24-hours or 1440 minutes, regression  
 3 coefficients can be interpreted on a meaningful scale.

4

5 The *ilr* multiple linear regression model for  $n$  compositional observations

6  $(\mathbf{x}_i, y_i)$ ,  $i = 1, 2, \dots, n$ , where  $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{iD}]$  with  $\sum_{j=1}^D x_{ij} = 1$ , is

7 
$$y_i = \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \varepsilon_i \quad (5)$$

8 where

9 
$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{x_{ij}}{\sqrt[3]{\prod_{k=j+1}^D x_{ik}}} \right) \text{ for } j = 1, 2, \dots, D-1,$$

10 with intercept  $\beta_0$ , regression parameters  $\beta_1, \beta_2, \dots, \beta_{D-1}$  and  $\varepsilon_i \sim N(0, \sigma^2)$  independently.

11 The regression coefficient  $\beta_1$  represents the change in the response variable when the

12 first *ilr* coordinate is changed while the remaining *ilr* coordinates are all kept constant

13 and the total sum is maintained, i.e.,  $\sum_{j=1}^D x_{ij} = 1$ . When the component in the

14 numerator of the first *ilr* coordinate is increased by one factor, all the components in the

15 denominator can simultaneously be decreased by another factor to maintain the constant

16 total. For example,  $ilr_1$  for a 4-part composition becomes

17 
$$\sqrt{\frac{3}{4}} \ln \left( \frac{x_1}{\sqrt[3]{(x_2 \cdot x_3 \cdot x_4)}} \cdot \frac{1+r}{1-s} \right). \quad (6)$$

1 The remaining *ilr* coordinates are then all kept constant, as they contain only the  
 2 components from the denominator of the first *ilr* coordinate which are all decreased by  
 3 the same proportion. Continuing with the example above, the remaining *ilr* coordinates,  
 4 i.e., *ilr*<sub>2</sub> and *ilr*<sub>3</sub>, respectively become

$$5 \quad \sqrt{\frac{2}{3}} \ln \left( \frac{x_2}{\sqrt{(x_3 \cdot x_4)}} \cdot \frac{1-s}{1-s} \right) \text{ and } \sqrt{\frac{1}{2}} \ln \left( \frac{x_3}{x_4} \cdot \frac{1-s}{1-s} \right). \quad (7)$$

6 The estimated change in an outcome ( $\Delta\hat{y}$ ) when the first component ( $x_1$ ) of a  
 7 composition of interest (e.g., the mean composition for a particular population group) is  
 8 multiplied by  $1+r$  (other coordinates are kept constant as each remaining compositional  
 9 part is simultaneously multiplied by  $1-s$ ) can be calculated as

$$10 \quad \Delta\hat{y} = \hat{\beta}_1 \cdot \sqrt{\frac{D-1}{D}} \cdot \ln \left( \frac{1+r}{1-s} \right) \quad (8)$$

$$11 \quad \text{where } -1 < r < \frac{1-x_1}{x_1} \text{ and } s = r \cdot \frac{x_1}{1-x_1}.$$

12 The derivation of Equation (8) above, the corresponding  $100(1-\alpha)\%$  confidence interval  
 13 and the effect of including additional predictor variables into the *ilr* multiple linear  
 14 regression model is provided in Supplementary file 1.

15

16 To facilitate the translation of research findings to clinical practice, it is of interest to  
 17 estimate the health effects related to a meaningful quantum change (in minutes or hours)  
 18 of one part of the activity behaviour composition (relative to compensatory change – to

1 maintain a total of 24 hours – of the geometric mean of the remaining compositional  
2 parts). The next section illustrates a novel, meaningful interpretation of *ilr* beta  
3 coefficients from regression analysis of a specific type of data: compositional, and  
4 constrained to a meaningful constant sum (24 h or 1440 min), using an epidemiological  
5 dataset as an example.

6

#### 7 **4 Example: daily activity and adiposity**

8 The examples use data from the International Study of Childhood Obesity, Lifestyle and  
9 the Environment (ISCOLE), a large international study of children aged 9-11 years,  
10 conducted between 2011 and 2013.<sup>24</sup> Children were from urban and suburban centres in  
11 12 countries (Australia, Brazil, Canada, China, Colombia, Finland, India, Kenya,  
12 Portugal, South Africa, England, and the United States). Ethical approval for ISCOLE  
13 was obtained from the Institutional Review Board of the Pennington Biomedical  
14 Research Center in Baton Rouge, Louisiana, USA, and site-specific ethical approval  
15 was also received at each participating study site. Parental written informed consent and  
16 child assent were obtained as required by local review boards. Daily activity was  
17 measured by 7-day 24-hour accelerometry.<sup>25</sup> Nocturnal sleep duration was estimated  
18 using a fully automated algorithm.<sup>26, 27</sup> Once total sleep time and awake non-wear time  
19 (any sequence of  $\geq 20$  consecutive minutes of 0 activity counts) were removed, data  
20 were processed in 15-s epochs to determine sedentary time ( $\leq 25$  counts per 15 s), light

1 PA (26-573 counts per 15 s), and MVPA ( $\geq 574$  counts per 15 s), congruent with  
2 Evenson's cut-points.<sup>28</sup> For analysis, each component of 24-h time use (sleep, sedentary  
3 time, light PA and MVPA) was accumulated, weighted for weekdays:weekend days at  
4 5:2. No zero values were present among the compositional daily activity behaviour data.  
5 Adiposity was represented by body mass index (BMI) from measured weight and height  
6 ( $\text{BMI} = \text{weight [kg]} / \text{height [m}^2\text{]}$ ). Note that BMI is a positive real random variable with  
7 an absolute scale. Participant BMI was converted to z-scores using age- and sex-specific  
8 World Health Organization (WHO) reference data.<sup>29</sup> After its transformation, zBMI can  
9 take on positive and negative values, satisfying the assumption that the response  
10 variable has support on the real line. The analyses included 5828 children (2633 boys,  
11 3195 girls), with mean zBMI= 0.45 (SD=1.26).

12

13 First, the relationship between zBMI and the four-part daily activity composition (sleep,  
14 sedentary time, light PA, MVPA) was examined. Table 1 shows the results of the  
15 compositional multiple linear regression models, which included the *ilr* coordinates and  
16 terms for sex, highest parental education, number of parents, number of siblings, and  
17 study site. The *ilr* coordinates were calculated using (3) in which  $\boldsymbol{x}$  represented  
18 proportions of time spent in sleep, sedentary time, light PA and MVPA. Four sets of *ilr*-  
19 coordinate systems were constructed, each time rotating the sequence of activity  
20 behaviours, so that each behaviour was iteratively represented as the first compositional

1 part. Each of these *ilr*-coordinate systems was used in a multiple linear regression  
 2 model, where the regression coefficient of the first *ilr*-coordinate contained all the  
 3 information regarding the first activity component, relative to all the remaining  
 4 components. Therefore, only the regression coefficients corresponding to the first *ilr*-  
 5 coordinates are displayed in Table 1.

6  
 7 **Table 1.** Multiple linear regression analyses of the relationship between first isometric  
 8 log-ratio (*ilr*) coordinates and Body Mass Index (BMI) z-scores (Compositional  
 9 models).

<i>ilr</i> regression models	$\hat{\beta}$	SE	<i>t</i> -value	<i>p</i>
Model 1: $ilr_1 \propto \ln(\text{Sleep :}$ geometric mean of remaining behaviours)	-0.82	0.13	-6.22	<0.001
Model 2: $ilr_1 \propto \ln(\text{Sedentary time :}$ geometric mean of remaining behaviours)	0.35	0.10	3.30	<0.001
Model 3: $ilr_1 \propto \ln(\text{Light PA :}$	1.34	0.10	13.19	<0.001

geometric mean of remaining behaviours)

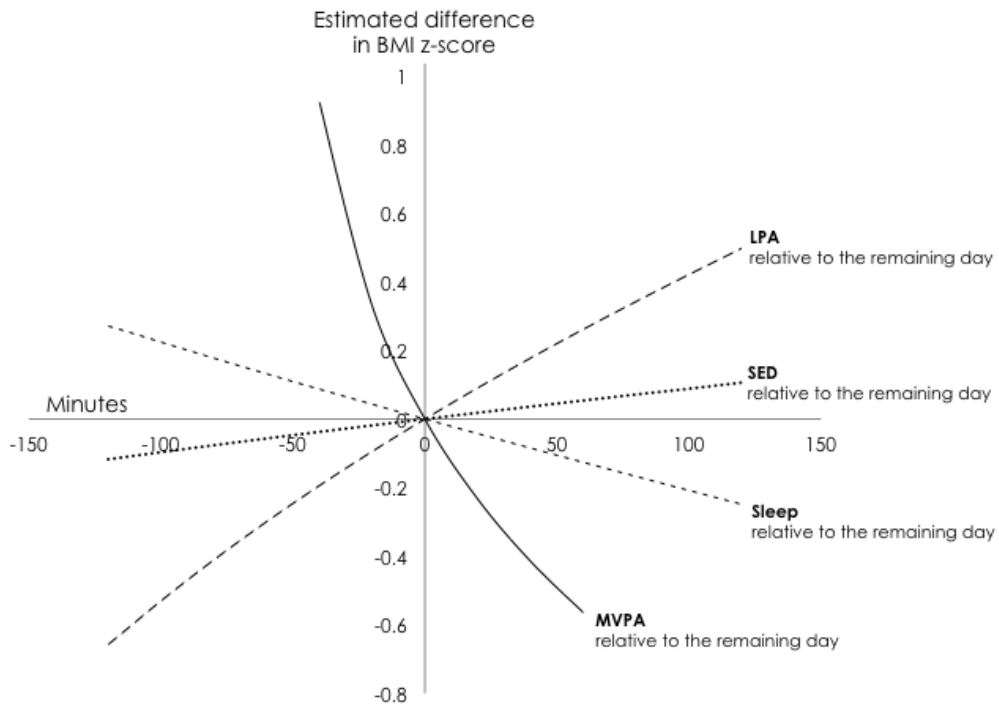
Model 4:  $ilr_1 \propto \ln(\text{MVPA} :$

geometric mean of remaining behaviours)	-0.87	0.05	-16.11	<0.001
---	-------	------	--------	--------

---

1 *ilr*: Isometric log ratio, BMI z-score: Body Mass Index transformed to z-score using  
2 age- and sex-specific World Health Organization (WHO) reference data,  $\hat{\beta}$ :  
3 unstandardized regression coefficient estimate, SE: Standard error, Light PA: Light-  
4 intensity physical activity; MVPA: Moderate-to-vigorous-intensity physical activity.  
5 All models adjusted for sex, highest parental education level, number of siblings,  
6 number of parents and study site. Adjusted *R*-squared = 0.11.  
7  
8 Adiposity (zBMI) was positively related to the relative time spent in sedentary time and  
9 light PA, and negatively related to the relative time spent in sleep and MVPA. Figure 1  
10 shows the association effect-size from the linear models further interpreted to  
11 meaningful parameters. To create the plot, the estimated differences in zBMI related to  
12 difference in one activity relative to the remaining activities (with the mean activity  
13 behaviour composition of the sample as the reference, or starting composition) were  
14 calculated using (8) for 15 min time interval increments ranging from 0 to 60 min

1 (MVPA) or 0 to 120 minutes (sleep, sedentary time, light PA). Worked examples are  
2 presented in Supplementary file 2. The R functions to estimate differences in zBMI are  
3 available freely at <https://github.com/tystan/deltacomp>. From Figure 1, it can be seen  
4 that estimated zBMI is lower by about 0.33 units with 30 min higher relative MVPA,  
5 and zBMI is higher by about 0.60 units with 30 min lower relative MVPA, compared to  
6 zBMI at the mean composition. The non-linear/non-symmetrical nature of the estimated  
7 zBMI response curves can be seen in Figure 1. The figure can also be used to assess  
8 equivalence of activity behaviours in the relationship with zBMI; for example, a  
9 horizontal line drawn at a zBMI -0.1 shows that this estimated difference in zBMI is  
10 associated with either 48 min more sleep (relative to the remaining behaviours), 108  
11 min less sedentary time (relative to the remaining behaviours), 21 min less light PA  
12 (relative to the remaining behaviours) or 8 min more MVPA (relative to the remaining  
13 behaviours) than the mean activity behaviour composition. The minute values can be  
14 calculated from the linear models, as detailed in Supplementary file 2.  
15



1

2 **Figure 1.** The Relationship Between Daily Behaviours and zBMI, Estimated by

3 Compositional Linear Regression Models.

4 BMI: Body Mass Index; SED: Sedentary Time; LPA: Light-Intensity Physical Activity;

5 MVPA: Moderate-to-Vigorous-Intensity Physical Activity.

6 Difference in Minutes Modelled Around the Population Mean Composition of

7 (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45.

8



1 The relationship between zBMI and the activity behaviour composition was further  
 2 investigated using an SBP approach, with *ilr* coordinates constructed according to  
 3 procedures outlined in Egozcue & Pawlowsky-Glahn.<sup>16</sup> The first partition separated the  
 4 behaviour components into two groups; inactivity-related behaviours (sleep and  
 5 sedentary time), and activity-related behaviours (light PA and MVPA). The second  
 6 partition was between sleep and sedentary time, and the final partition was between  
 7 light PA and MVPA. Table 2 presents the results from the *ilr* log-ratio multiple linear  
 8 regression model based on this SBP.

9

10 **Table 2.** Multiple linear regression analysis of the relationship between the isometric  
 11 log-ratio (*ilr*) coordinates obtained from a sequential binary partition and Body Mass  
 12 Index (BMI) z-scores (Compositional model).

	$\hat{\beta}$	SE	<i>t</i> -value	<i>p</i>
<hr/>				
<i>ilr</i> <sub>1</sub> ∝ ln(geometric mean of Sleep & Sedentary time : geometric mean of Light PA & MVPA)	-0.41	0.08	-5.33	<0.001
<i>ilr</i> <sub>2</sub> ∝ ln(Sleep : Sedentary time)	-0.72	0.13	-5.29	<0.001

$$ilr_3 \propto \ln(\text{Light PA} : \text{MVPA}) \quad -1.35 \quad 0.08 \quad -16.26 \quad <0.001$$

1 *ilr*: Isometric log ratio, BMI z-score: Body Mass Index transformed to z-score using  
2 age- and sex-specific World Health Organization (WHO) reference data,  $\hat{\beta}$ :  
3 unstandardized regression coefficient estimate, SE: Standard error, Light PA: Light-  
4 intensity physical activity, MVPA: Moderate-to-vigorous intensity physical activity.  
5 Model adjusted for sex, highest parental education level, number of siblings, number of  
6 parents and study site. Adjusted *R*-squared = 0.11.

7

8 Regression parameters suggest that, with increase in inactivity-related behaviours  
9 relative to decrease in activity-related behaviours, zBMI decreases (Table 2). This  
10 finding is explained by the co-consideration of light PA and MVPA as activity-related  
11 behaviours, i.e., the denominator of the first *ilr* coordinate is the geometric mean of both  
12 light PA and MVPA. As shown in the previous analysis (Table 1), light PA (relative to  
13 all remaining behaviours) has a strong positive association with zBMI. Predicted  
14 increase in zBMI for increase in light PA (accompanied by corresponding decrease in  
15 all other behaviours) was higher than the respective effect sizes for any other behaviour  
16 (relative to the remaining behaviours) (Figure 1). This is because whatever light PA  
17 replaces (sleep, MVPA, and more surprisingly, sedentary time) is associated with lower  
18 fatness. The regression coefficient for the second *ilr* coordinate from the SBP (Table 2)  
19 implies that the increase in sleep relative to sedentary time is associated with lower

1 expected zBMI. The third *ilr* regression coefficient indicates that an increase in light PA  
2 at the expense of MVPA is associated with higher expected zBMI. The results obtained  
3 from the regression model from the SBP are consistent and complementary to the  
4 results from the previous *ilr* models. With the SBP we obtain complementary  
5 information about the substitution of time between selected groups of parts. Of course,  
6 with an SBP as presented here, none of the coordinates extract all the relative  
7 information about any of the behaviours. This might be an advantage because of the  
8 danger that the geometric mean of the other components may itself conceal some  
9 unpredictable patterns and potentially affect the interpretability of the first coordinate.  
10 On the other hand, the concrete choice of SBP necessarily entails a substantial amount  
11 of subjectivity and, additionally, the effects of merging information contained in  
12 compositional parts by taking the geometric mean will also influence the interpretation  
13 of the first coordinate.

14

15 To further illustrate the importance of using a compositional approach, we analyzed the  
16 same dataset using standard (non-compositional) regression (Table 3). The purpose of  
17 ensuing analysis is not to compare the findings of standard regression models to  
18 findings of compositional regression models, but to demonstrate the potential pitfalls  
19 and limitations of standard multiple linear regression when analyzing data of a  
20 compositional nature. Raw values of time spent in each behaviour (min/day) were used,

1 with zBMI as the dependent variable. It is not possible to include all daily activity  
 2 behaviours (sleep, sedentary time, light PA and MVPA) in the regression model as this  
 3 would result in a singular covariance matrix. Therefore four models were used, with  
 4 each model iteratively excluding a different behaviour, i.e., (1) excluded sleep; (2)  
 5 excluded sedentary time; (3) excluded light PA; (4) excluded MVPA. The four analyzed  
 6 models represent the most adjusted traditional regression models. Variance Inflation  
 7 Factors (VIF) for Models 1 - 3 ranged between 1.9 and 2.4, indicating multi-collinearity  
 8 was likely not a concern.<sup>30</sup> However, VIF for model 4 were high (between 7.2 and  
 9 11.3), suggesting potential instability of regression estimates. Each model was  
 10 additionally adjusted for sex, highest parental education, number of parents, number of  
 11 siblings, and study site.

12

13 **Table 3.** Traditional multiple linear regression analysis of the relationship  
 14 between raw daily activity data (min/day) and Body Mass Index (BMI) z-scores  
 15 (Non-compositional models).

	$\hat{\beta}$	SE	<i>t</i> -value	<i>p</i>
Model 1				
(Sleep excluded)				
Sedentary time	0.002	0.000	5.27	<0.001
Light PA	0.005	0.000	11.48	<0.001

---

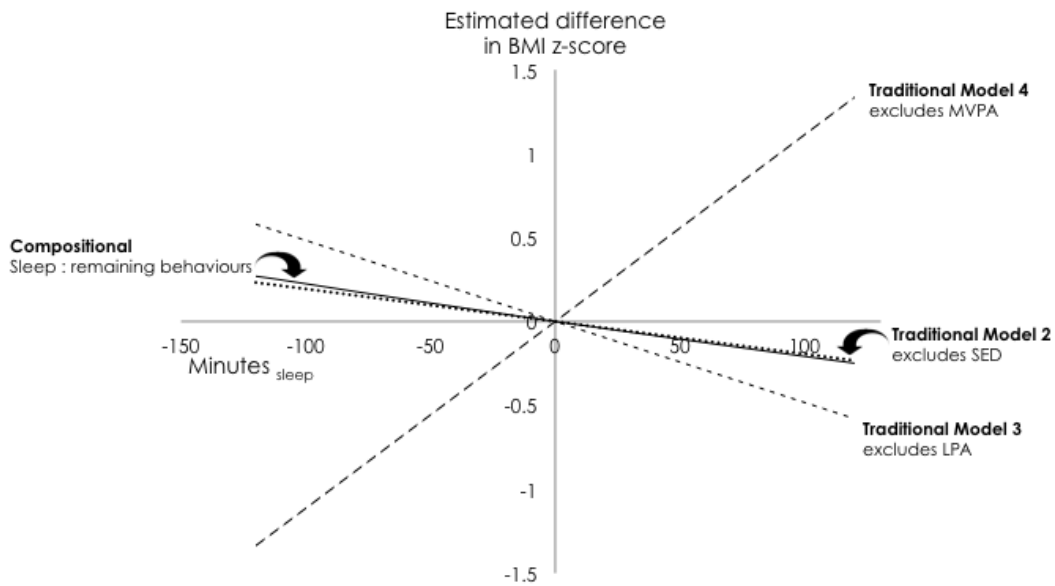
MVPA	-0.011	0.001	-13.17	<0.001
Model 2				
Sleep	-0.002	0.000	-5.27	<0.001
(Sedentary time excluded)				
Light PA	0.003	0.000	8.33	<0.001
MVPA	-0.013	0.001	-17.32	<0.001
Model 3				
Sleep	-0.005	0.000	-11.48	<0.001
Sedentary time	-0.003	0.000	-8.33	<0.001
(Light PA excluded)				
MVPA	-0.016	0.001	-17.35	<0.001
Model 4				
Sleep	0.011	0.001	13.17	<0.001
Sedentary time	0.013	0.001	17.32	<0.001

---

Light PA	0.016	0.001	17.35	<0.001
(MVPA excluded)				

1 BMI z-score: Body Mass Index transformed to z-score using age- and sex-specific  
2 World Health Organization (WHO) reference data,  $\hat{\beta}$ : unstandardized regression  
3 coefficient estimate, SE: Standard error, Light PA: Light-intensity physical activity,  
4 MVPA: Moderate-to-vigorous-intensity physical activity.  
5 All models adjusted for sex, highest parental education level, number of  
6 siblings, number of parents and study site. Adjusted *R*-squared for all models =  
7 0.11.  
8  
9 The regression estimates from the traditional regression analysis were inconsistent  
10 across the models. This demonstrates that the choice of omitted behaviour may have  
11 substantial influence on the interpretation of the relationships between the remaining  
12 behaviours and zBMI. Moreover, the regression coefficients for sleep and sedentary  
13 time varied from positive to negative, depending on the model. Figures 2 - 5 depict the  
14 inconsistency of the regression coefficients across traditional regression models and  
15 how they differ from the regression coefficients obtained from compositional models.  
16 The inconsistent findings from traditional regression models demonstrate that results  
17 from traditional analyses are unreliable when raw untransformed minutes are used as  
18 activity behaviour inputs.

1



2

3 **Figure 2:** The Relationship Between Sleep and zBMI: Comparison between

4 Compositional and Traditional Regression Models.

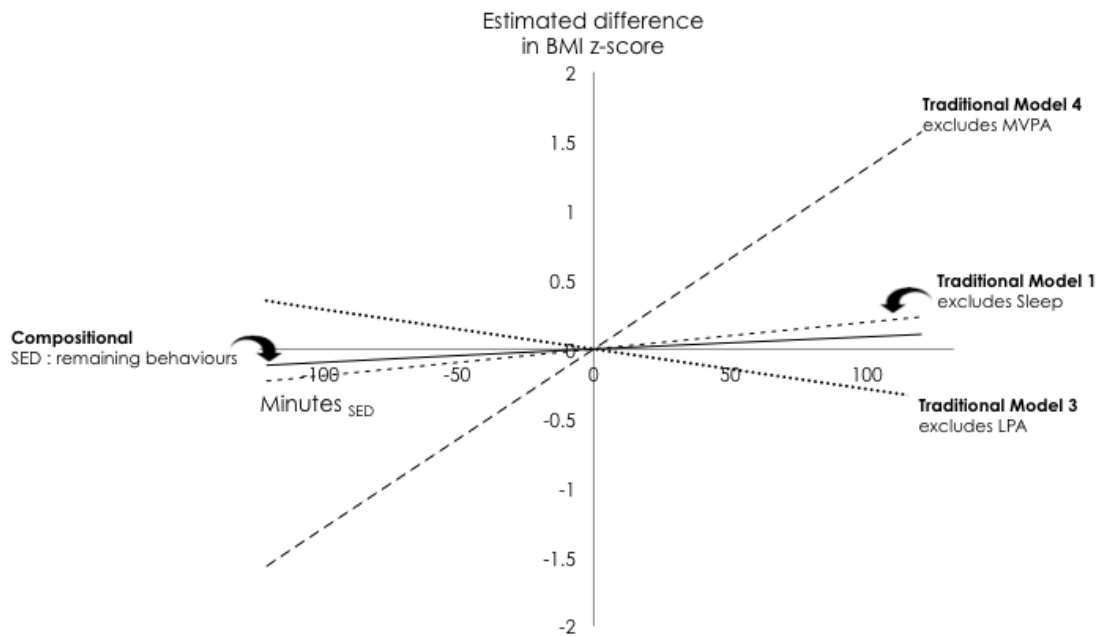
5 SED: Sedentary Time; LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-

6 Vigorous-Intensity Physical Activity.

7 Difference in Minutes Modelled Around the Population Mean Composition of

8 (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45.

9



1

2

3 **Figure 3.** The Relationship Between Sedentary Time and zBMI: Comparison between

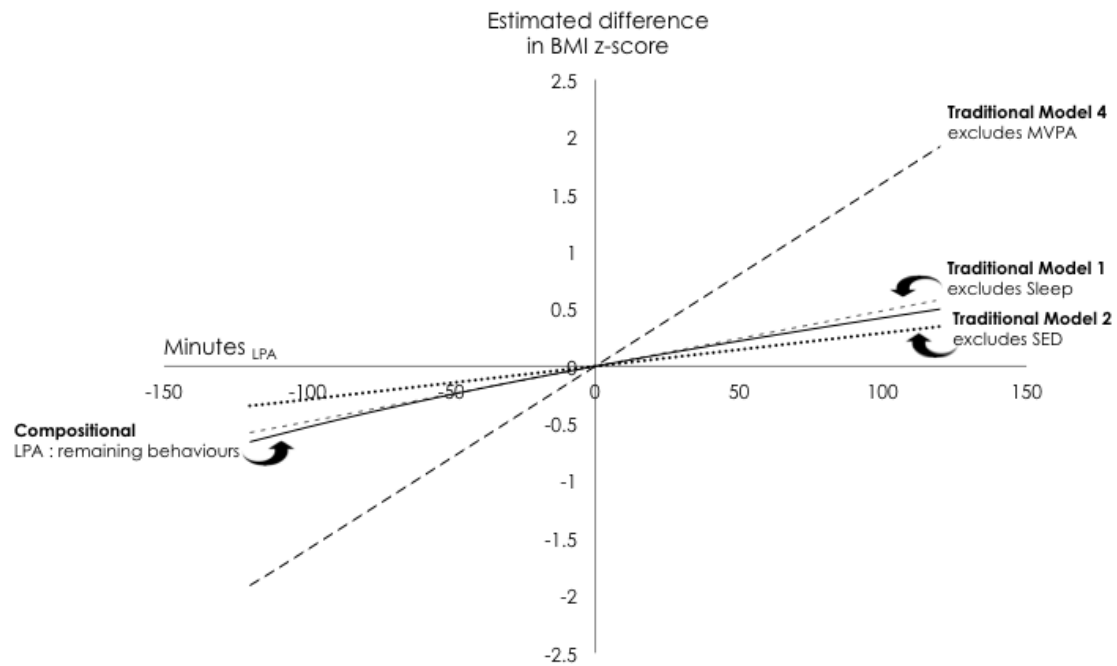
4 Compositional and Traditional Regression Models. SED: Sedentary Time; LPA: LPA:

5 Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical

6 Activity. Difference in Minutes Modelled Around the Population Mean Composition of

7 (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45

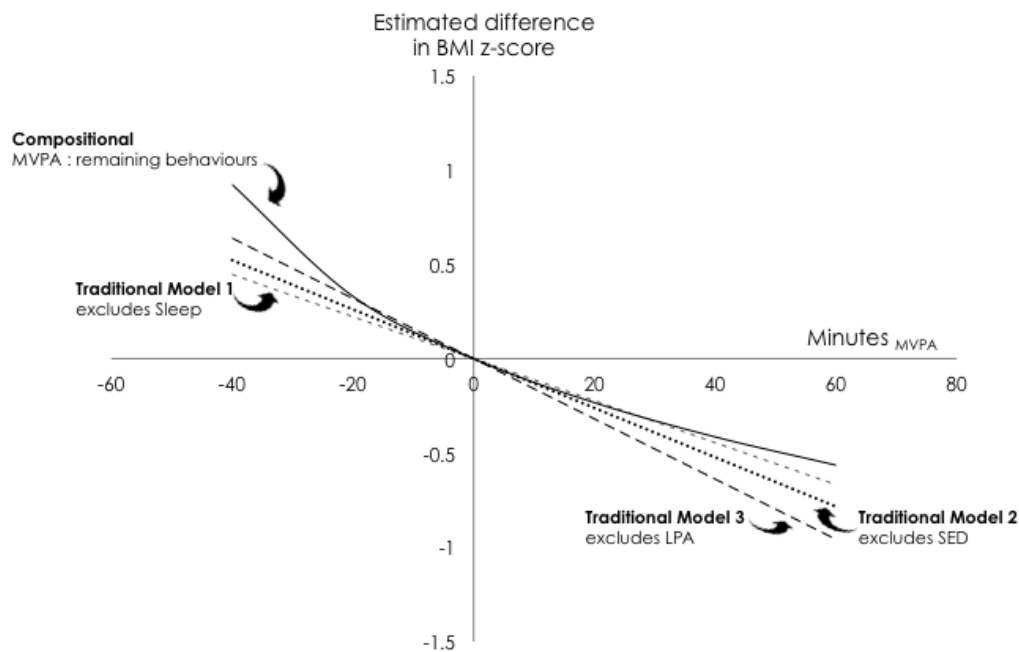




1

2 **Figure 4.** The Relationship Between Light Physical Activity and zBMI: Comparison  
 3 between Compositional and Traditional Regression Models. SED: Sedentary Time;  
 4 LPA: Light-Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity  
 5 Physical Activity. Difference in Minutes Modelled Around the Population Mean  
 6 Composition of (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean  
 7 zBMI of 0.45.

8



1

2 **Figure 5.** The Relationship Between MVPA and zBMI: Comparison between  
 3 Compositional and Traditional Regression Models. SED: Sedentary Time; LPA: Light-  
 4 Intensity Physical Activity; MVPA: Moderate-to-Vigorous-Intensity Physical Activity.  
 5 Difference in Minutes Modelled Around the Population Mean Composition of  
 6 (min/day): Sleep=539; SED=525; LPA=320; MVPA=57, and Mean zBMI of 0.45

7

8 **5 Comments**

1 While the statistical issue presented by the singularity of daily activity data has been  
2 acknowledged in previous literature,<sup>31-33</sup> there has been little consensus on how this  
3 might be addressed. In fact, previous research has overwhelmingly overlooked the  
4 compositional nature of time use data, and has considered daily behaviours as  
5 individual, absolute quantities. Traditional regression does not account for the closed  
6 nature of time use and the ensuing co-dependence of daily behaviours, and,  
7 consequently, findings may be spurious.<sup>8, 13</sup> The incontestable inability to account for  
8 all activity behaviours due to their perfect multi-collinearity is the main limitation of the  
9 traditional (non-compositional) regression models. Omitting one or more behaviours to  
10 be able to run a traditional regression analysis has been a widely used approach that  
11 ignores the true compositional nature of the data and seems to result in inconsistent  
12 regression estimates. The potential for erroneous results from traditional regression was  
13 demonstrated in this study, with estimates for sleep and sedentary time indicating a  
14 positive relationship with zBMI in some models and an inverse relationship with zBMI  
15 in other models. Such inconsistent results have the potential to undermine the credibility  
16 of academic and public health messages. For example, findings from traditional Models  
17 1 and 4 indicate public policy should focus on the reduction of sedentary time in  
18 children, whilst findings from Model 3 indicate higher sedentary time should be  
19 encouraged to potentially reduce adiposity. Model checks (VIF) detected multi-  
20 collinearity issues for Model 4, however, VIF values for the remaining models were

1 below acceptable thresholds and were therefore unable to explain the inconsistencies  
2 between Model 1 and 3. This suggests that VIF is not an acceptable diagnostic indicator  
3 for compositional data. Uncertainty regarding the role of sedentary time is commonly  
4 observed in contemporary health research, with some studies finding strong associations  
5 with adiposity,<sup>34-37</sup> and other studies finding none.<sup>38-41</sup> Contemporary confusion  
6 regarding the relationship between activity behaviours and health may be a consequence  
7 of a flawed approach to statistical analysis.

8

9 Findings from the compositional data analysis (based on *ilr* coordinates) applied in this  
10 study correspond to results obtained from some traditional regression models used in  
11 this study and also some evidence from previous research.<sup>42-45</sup> However, because  
12 traditional models are unable to adjust for all remaining daily behaviours, outcomes  
13 may be unreliable, and are not directly comparable to outcomes from the compositional  
14 models. Furthermore, unlike compositional regression, traditional regression models are  
15 unable to detect asymmetry in associations depending on whether a behaviour is  
16 increased or decreased, or whether associations differ at various daily time-use  
17 compositions. Traditional regression can therefore not discern between the importance  
18 of maintaining or increasing a behaviour, or whether change in behaviour has a stronger  
19 association with adiposity for children with differing time-use compositions (e.g. active

1 children compared with sedentary children). Such considerations are important for  
2 informing public health messages and potential intervention strategies.

3  
4 This study's intention was primarily to demonstrate compositional data analysis of daily  
5 activity data and interpret the findings in a meaningful manner. Therefore, analyses  
6 were carried out on the complete international dataset, however, it must be remembered  
7 that the data used were cross-sectional, and therefore causality cannot be inferred. To  
8 further investigate the relationship between activity and health, and to guide the  
9 planning of interventions and public health policy, future compositional data analyses  
10 should be performed on longitudinal data, and may be stratified by sex and be country-  
11 specific. In addition, compositional data analysis techniques for other statistical  
12 applications should be explored (e.g., isotemporal substitution, cluster analysis,  
13 principal component analysis), and include non-compositional lifestyle behaviours, such  
14 as diet quality.

15  
16 The inadequacy of traditional linear regression models for data of a closed, and  
17 therefore relative nature, is well acknowledged. However, in health research, the closed  
18 nature of daily time use has been largely ignored. This study demonstrates the potential  
19 for incorrect outcomes from models commonly applied in the field. The findings imply  
20 that previous studies, including evidence accepted as being high level such as

1 systematic reviews, which agglomerate such studies, may be erroneous and cannot be  
2 trusted. The implications of the application of compositional data analysis to health  
3 research are quite profound. Since almost all previous analyses of the associations  
4 between time use and health outcomes have used methods incompatible with  
5 compositional data, they are all to some extent vitiated, and should be interpreted with  
6 caution. A shift towards an integrated, compositional data analysis approach, where all  
7 daily activity behaviours are considered, should be a priority. Not until robust research  
8 methodologies are implemented can valid estimates of the health associations of daily  
9 activity behaviours be made, and the mortality and economic impact of sleep, sedentary  
10 time, and physical inactivity be assessed.

11

#### 12 **Funding acknowledgements and declaration of conflicting interests**

13 This work was supported by an Australian Government Research Training Program  
14 Scholarship [DD], the National Heart Foundation (100188) [CM] and partially  
15 supported by the Spanish Ministry of Economy and Competitiveness under the project  
16 CODA-RETOS (MTM2015-65016- C2-1(2)-R) [JAMF]. The example dataset used in  
17 this study is from the International Study of Childhood Obesity, Lifestyle and  
18 Environment (ISCOLE), which was funded by The Coca-Cola Company. The funders  
19 had no role in the design and conduct of the study, data collection, management,  
20 analysis and interpretation of the data; and decision to publish, preparation, review or

1 approval of this manuscript. DD, TES, JAMF and ZP are not part of ISCOLE group.  
2 The authors declare that they have no competing interests. The authors gratefully  
3 acknowledge the constructive comments of the referees which have undoubtedly helped  
4 to improve the quality of the paper.

5

6

7 **Data availability statement**

8 The data that support the findings of this study are available from Peter T. Katzmarzyk  
9 ([Peter.Katzmarzyk@pbrc.edu](mailto:Peter.Katzmarzyk@pbrc.edu)) but restrictions apply to the availability of these data,  
10 which were used under license for the current study, and so are not publicly available.  
11 Data are however available from the authors upon reasonable request and with  
12 permission of Pennington Biomedical Research Center.

13

## 1 **References**

- 2 1. Lee I-M, Shiroma EJ, Lobelo F, et al. Effect of physical inactivity on major non-  
3 communicable diseases worldwide: an analysis of burden of disease and life  
4 expectancy. *The Lancet*. 2012; 380: 219-29.
- 5 2. Ding D, Lawson KD, Kolbe-Alexander TL, et al. The economic burden of  
6 physical inactivity: a global analysis of major non-communicable diseases. *The Lancet*.  
7 2016; 388: 1311-24.
- 8 3. Hirshkowitz M, Whiton K, Albert SM, et al. National Sleep Foundation's sleep  
9 time duration recommendations: methodology and results summary. *Sleep Health*.  
10 2015; 1: 40-3.
- 11 4. Wilmot EG, Edwardson CL, Achana FA, et al. Sedentary time in adults and the  
12 association with diabetes, cardiovascular disease and death: systematic review and  
13 meta-analysis. *Diabetologia*. 2012; 55: 2895-905.
- 14 5. Chaput J-P, Gray CE, Poitras VJ, et al. Systematic review of the relationships  
15 between sleep duration and health indicators in school-aged children and youth 1. *Appl*  
16 *Physiol Nutr Metab*. 2016; 41: S266-S82.
- 17 6. Carson V, Hunter S, Kuzik N, et al. Systematic review of sedentary behaviour  
18 and health indicators in school-aged children and youth: an update 1. *Appl Physiol Nutr*  
19 *Metab*. 2016; 41: S240-S65.



- 1 7. Rezende LFM, Sá TH, Mielke GI, et al. All-Cause Mortality Attributable to  
2 Sitting Time: Analysis of 54 Countries Worldwide. *Am J Prev Med.* 2016; 51: 253-63.
- 3 8. Pedišić Ž. Measurement issues and poor adjustments for physical activity and  
4 sleep undermine sedentary behaviour research—the focus should shift to the balance  
5 between sleep, sedentary behaviour, standing and activity. *Kinesiology.* 2014; 46: 135-  
6 46.
- 7 9. Aitchison J. The statistical analysis of compositional data. *Journal of the Royal*  
8 *Statistical Society Series B (Methodological).* 1982; 44: 139-77.
- 9 10. Barceló-Vidal C and Martín-Fernández J-A. The mathematics of compositional  
10 analysis. *Austr J Stat.* 2016; 45: 57-71.
- 11 11. Pawlowsky-Glahn V, Egozcue JJ and Tolosana-Delgado R. *Modeling and*  
12 *analysis of compositional data.* John Wiley & Sons, 2015.
- 13 12. Martín-Fernández J-A, Daunis-i-Estadella J and Mateu-Figueras G. On the  
14 interpretation of differences between groups for compositional data. *SORT-Statistics*  
15 *and Operations Research Transactions.* 2015; 39: 231-52.
- 16 13. Pearson K. Mathematical Contributions to the Theory of Evolution.--On a Form  
17 of Spurious Correlation Which May Arise When Indices Are Used in the Measurement  
18 of Organs. *Proceedings of the Royal Society of London.* 1896; 60: 489-98.
- 19 14. Mateu-Figueras G, Pawlowsky-Glahn V and Egozcue J. The Principle of  
20 Working on Coordinates. In: Pawlowsky-Glahn V and Buccianti A, (eds.).

- 1 *Compositional Data Analysis: Theory and Applications*. Chichester, UK: John Wiley &  
2 Sons, 2011, p. 29-42.
- 3 15. Palarea-Albaladejo J and Martín-Fernández JA. zCompositions—R package for  
4 multivariate imputation of left-censored data under a compositional approach.  
5 *Chemometr Intell Lab*. 2015; 143: 85-96.
- 6 16. Egozcue JJ and Pawlowsky-Glahn V. Groups of parts and their balances in  
7 compositional data analysis. *Math Geol*. 2005; 37: 795-828.
- 8 17. Hron K, Filzmoser P and Thompson K. Linear regression with compositional  
9 explanatory variables. *J Appl Stat*. 2012; 39: 1115-28.
- 10 18. Chastin SF, Palarea-Albaladejo J, Dontje ML, et al. Combined effects of time  
11 spent in physical activity, sedentary behaviors and sleep on obesity and cardio-  
12 metabolic health markers: A novel compositional data analysis approach. *PLoS ONE*.  
13 2015; 10: e0139984.
- 14 19. Leite MLC. Applying compositional data methodology to nutritional  
15 epidemiology. *Stat Methods Med Res*. 2016; 25: 3057-65.
- 16 20. Mert MC, Filzmoser P, Endel G, et al. Compositional data analysis in  
17 epidemiology. *Stat Methods Med Res*. Epub ahead of print 6 October 2016: DOI:  
18 10.1177/0962280216671536.
- 19 21. Tsilimigras MC and Fodor AA. Compositional data analysis of the microbiome:  
20 fundamentals, tools, and challenges. *Ann Epidemiol*. 2016; 26: 330-5.

- 1 22. Carson V, Tremblay MS, Chaput J-P, et al. Associations between sleep duration,  
2 sedentary time, physical activity, and health indicators among Canadian children and  
3 youth using compositional analyses 1. *Appl Physiol Nutr Metab.* 2016; 41: S294-S302.
- 4 23. Muller I, Hron K, Fiserova E, et al. Interpretation of Compositional Regression  
5 with Application to Time Budget Analysis. *arXiv preprint.* 2016: arXiv:1609.07887.
- 6 24. Katzmarzyk PT, Barreira TV, Broyles ST, et al. The international study of  
7 childhood obesity, lifestyle and the environment (ISCOLE): Design and methods. *BMC*  
8 *Public Health.* 2013; 13: E900.
- 9 25. Tudor-Locke C, Barreira TV, Schuna JM, et al. Improving wear time  
10 compliance with a 24-hour waist-worn accelerometer protocol in the International Study  
11 of Childhood Obesity, Lifestyle and the Environment (ISCOLE). *Int J Behav Nutr Phys*  
12 *Act.* 2015; 12: 1-9.
- 13 26. Tudor-Locke C, Barreira TV, Schuna Jr JM, et al. Fully automated waist-worn  
14 accelerometer algorithm for detecting children's sleep-period time separate from 24-h  
15 physical activity or sedentary behaviors. *Appl Physiol Nutr Metab.* 2013; 39: 53-7.
- 16 27. Barreira TV, Schuna Jr JM, Mire EF, et al. Identifying children's nocturnal sleep  
17 using 24-h waist accelerometry. *Med Sci Sports Exerc.* 2015; 47: 937-43.
- 18 28. Evenson KR, Catellier DJ, Gill K, et al. Calibration of two objective measures  
19 of physical activity for children. *J Sports Sci.* 2008; 26: 1557-65.

- 1 29. de Onis M, Onyango A, Borghi E, et al. Development of a WHO growth  
2 reference for school-aged children and adolescents. *Bull World Health Organ.* 2007.
- 3 30. Fox J and Monette G. Generalized collinearity diagnostics. *J Am Stat Assoc.*  
4 1992; 87: 178-83.
- 5 31. Augustin NH, Mattocks C, Cooper AR, et al. Modelling fat mass as a function  
6 of weekly physical activity profiles measured by Actigraph accelerometers. *Physiol*  
7 *Meas.* 2012; 33: 1831.
- 8 32. Hunt E, McKay E, Fitzgerald A, et al. Time use and daily activities of late  
9 adolescents in contemporary Ireland. *J Occup Sci.* 2014; 21: 42-64.
- 10 33. Maher C, Olds T, Mire E, et al. Reconsidering the sedentary behaviour  
11 paradigm. *PLoS ONE.* 2014; 9: e86403.
- 12 34. Hussey J, Bell C, Bennett K, et al. Relationship between the intensity of  
13 physical activity, inactivity, cardiorespiratory fitness and body composition in 7–10-  
14 year-old Dublin children. *Br J Sports Med.* 2007; 41: 311-6.
- 15 35. Kennedy K, Shepherd S, Williams JE, et al. Activity, body composition and  
16 bone health in children. *Arch Dis Child.* 2013; 98: 204-7.
- 17 36. Machado-Rodrigues AM, Coelho-e-Silva MJ, Mota J, et al. Cardiorespiratory  
18 fitness, weight status and objectively measured sedentary behaviour and physical  
19 activity in rural and urban Portuguese adolescents. *J Child Health Care.* 2012; 16: 166-  
20 77.

- 1 37. Marques A, Minderico C, Martins S, et al. Cross-sectional and prospective  
2 associations between moderate to vigorous physical activity and sedentary time with  
3 adiposity in children. *Int J Obes*. 2016; 40: 28-33.
- 4 38. Loprinzi PD, Cardinal BJ, Lee H, et al. Markers of adiposity among children  
5 and adolescents: implications of the isothermal substitution paradigm with sedentary  
6 behavior and physical activity patterns. *J Diabetes Metab Disord*. 2015; 14: 1.
- 7 39. Dowd KP, Harrington DM, Hannigan A, et al. Light-intensity physical activity  
8 is associated with adiposity in adolescent females. *Med Sci Sports Exerc*. 2014; 46:  
9 2295-300.
- 10 40. Basterfield L, Pearce MS, Adamson AJ, et al. Physical activity, sedentary  
11 behavior, and adiposity in English children. *Am J Prev Med*. 2012; 42: 445-51.
- 12 41. Hjorth MF, Chaput J-P, Damsgaard CT, et al. Low physical activity level and  
13 short sleep duration are associated with an increased cardio-metabolic risk profile: a  
14 longitudinal study in 8-11 year old Danish children. *PLoS ONE*. 2014; 9: e104677.
- 15 42. Saunders TJ, Gray CE, Poitras VJ, et al. Combinations of physical activity,  
16 sedentary behaviour and sleep: relationships with health indicators in school-aged  
17 children and youth 1. *Appl Physiol Nutr Metab*. 2016; 41: S283-S93.
- 18 43. Tremblay MS, LeBlanc AG, Kho ME, et al. Systematic review of sedentary  
19 behaviour and health indicators in school-aged children and youth. *Int J Behav Nutr*  
20 *Phys Act*. 2011; 8: 98.

1 44. Janssen I and LeBlanc AG. Review Systematic review of the health benefits of  
2 physical activity and fitness in school-aged children and youth. *Int J Behav Nutr Phys*  
3 *Act.* 2010; 7: 1-16.

4 45. Chen X, Beydoun MA and Wang Y. Is sleep duration associated with childhood  
5 obesity? A systematic review and meta - analysis. *Obesity.* 2008; 16: 265-74.

6

7

1 **Supplementary file 1**

2

3 **The model**

4

5 As outlined in the main paper, the *ilr* model for  $n$  compositional observations

6  $(\mathbf{x}_i, y_i), i = 1, 2, \dots, n$ , where  $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{iD})^T$  with  $\sum_{j=1}^D x_{ij} = 1$ , is

7

8 
$$y_i = \beta_0 + \mathbf{z}_i^T \boldsymbol{\beta} + \epsilon_i$$

9

10 
$$= \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \epsilon_i, \tag{1}$$

11 where

12

13 
$$z_{ij} = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{x_{ij}}{(\prod_{k=j+1}^D x_{ik})^{1/(D-j)}} \right) \text{ for } j = 1, 2, \dots, D-1.$$

14

15

16 Note that covariates can be included without transform (subject to the standard

17 multiple linear regression assumptions) into the *ilr* multiple linear regression

18 model without alterations to the derived formula in Equation (8) of the main

19 paper. For completeness, should there be  $E$  covariates for each observation  $i$

1  $(w_{i1}, w_{i2}, \dots, w_{iE})$  to be added to the *ilr* multiple linear regression model, the model  
2 would be similarly specified as follows,

3

$$4 \quad y_i = \beta_0 + \sum_{j=1}^{D-1} \beta_j z_{ij} + \sum_{j=1}^E \beta_{D-1+j} w_{ij} + \epsilon_i.$$

5

### 6 **Relative changes in the components of $x_i$**

7 Consider a relative increase in  $x_{i1}$  by a factor of  $1 + r$  with  $-1 < r < \frac{1-x_1}{x_{i1}}$ . Note  
8 that  $r$  cannot take a value of  $\frac{1-x_1}{x_{i1}}$  (or greater) as

9

$$10 \quad (1 + r)x_{i1} < 1$$

11

$$12 \quad \Rightarrow rx_{i1} < 1 - x_{i1}$$

13

$$14 \quad \Rightarrow r < \frac{1-x_1}{x_{i1}}$$

15

16 To maintain the compositional components' sum to unity when a relative increase  
17 in  $x_{i1}$  by a factor of  $1 + r$  is applied, the remaining components can be reduced  
18 using a factor of  $1 - s$ . We derive the value of  $s$  below.



1

2 The individual components are adjusted as per below:

3

4

$$x_{i1} \rightarrow (1+r)x_{i1} = x_{i1}^*$$

5

6

$$x_{i2} \rightarrow (1+r)x_{i2} = x_{i2}^*$$

7

$$\vdots \quad \quad \quad \vdots \quad \quad \quad \vdots$$

8

$$x_{iD} \rightarrow (1+r)x_{iD} = x_{iD}^*$$

9

10 Note that  $\sum_{j=1}^D x_{ij} = 1$  and  $\sum_{j=1}^D x_{ij}^* = 1$ . Also note therefore  $\sum_{j=2}^D x_{ij} = 1 - x_{i1}$ .

11 Now to find an expression for  $s$ , consider the following:

12

13

$$\sum_{j=1}^D x_{ij}^* = 1$$

14

15

$$\Rightarrow (1+r)x_{i1} + \sum_{j=2}^D (1-s)x_{ij} = 1$$

16

17

$$\Rightarrow (1+r)x_{i1} + (1-s) \sum_{j=2}^D x_{ij} = 1$$

1

2

$$\Rightarrow (1+r)x_{i1} + (1-s)(1-x_{i1}) = 1$$

3

4

$$\Rightarrow rx_{i1} = s(1-x_{i1})$$

5

6

$$\Rightarrow s = r \frac{x_{i1}}{1-x_{i1}}$$

7

8

9

**Estimated change in the outcome for an increase in the first compositional**

10

**part and equal relative reductions in the remaining components**

11

12

Consider a new observation

13

14

$$\mathbf{x}_0 = (x_{01}, x_{02}, \dots, x_{0D})^T$$

15

16

and the corresponding *ilr* coordinates

17

18

$$\mathbf{z}_0 = (z_{01}, z_{02}, \dots, z_{0(D-1)})^T.$$

19

1 Now let us consider a new set of predictor variables

2

3

$$\mathbf{x}_0^* = ((1+r)x_{01}, (1-s)x_{02}, \dots, (1-s)x_{0D})^T$$

4

5 and the corresponding *ilr* coordinates

6

7

$$\mathbf{z}_0^* = (z_{01}^*, z_{02}^*, \dots, z_{0(D-1)}^*)^T$$

8

9 where

10

$$11 \quad z_{01}^* = \sqrt{\frac{D-1}{D}} \ln \left( \frac{x_{01}^*}{(\prod_{k=2}^D x_{0k}^*)^{1/(D-1)}} \right)$$

12

$$13 \quad = \sqrt{\frac{D-1}{D}} \ln \left( \frac{(1+r)x_{01}}{\left( ((1-s)^{D-1} (\prod_{k=2}^D x_{0k}) \right)^{1/(D-1)}} \right)$$

14

$$15 \quad = \sqrt{\frac{D-1}{D}} \ln \left( \frac{(1+r)}{(1-s)} \frac{x_{01}}{(\prod_{k=2}^D x_{0k})^{1/(D-1)}} \right), \text{ and}$$

16

$$1 \quad z_{0j}^* = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{x_{0j}^*}{\left( \prod_{k=j+1}^D x_{0k}^* \right)^{1/(D-1)}} \right) \text{ for } j = 2, 3, \dots, D-1$$

2

$$3 \quad = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{(1-s)x_{0j}}{\left( ((1-s)^{D-j} \prod_{k=j+1}^D x_{0k} \right)^{1/(D-j)}} \right)$$

4

$$5 \quad = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{(1-s)}{(1-s)} \frac{x_{0j}}{\left( \prod_{k=j+1}^D x_{0k} \right)^{1/(D-j)}} \right)$$

6

$$7 \quad = \sqrt{\frac{D-j}{D-j+1}} \ln \left( \frac{x_{0j}}{\left( \prod_{k=j+1}^D x_{0k} \right)^{1/(D-j)}} \right)$$

$$8 \quad = z_{01}.$$

9

10 That is, there is no change in the *ilr* coordinates for  $j = 2, 3, \dots, D-1$  and

11

$$12 \quad \mathbf{z}_0^* = (z_{01}^*, z_{02}^*, \dots, z_{0(D-1)}^*)^T.$$

13

14 The estimated outcome for predictors  $\mathbf{x}_0$  is

15

1 
$$\hat{y}_0 = \hat{\beta}_0 + z_0^T \hat{\beta}$$

2

3 and the estimated outcome for predictors  $\mathbf{x}_0^*$  is

4

5 
$$\hat{y}_0^* = \hat{\beta}_0 + z_0^{*T} \hat{\beta}.$$

6

7 Therefore the estimated change in the predicted outcome going from predictors  $\mathbf{x}_0$

8 to  $\mathbf{x}_0^*$  is

9

10 
$$\Delta \hat{y} = \hat{y}_0^* - \hat{y}_0$$

11

12 
$$= \hat{\beta}_0 + z_0^{*T} \hat{\beta} - \hat{\beta}_0 + z_0^T \hat{\beta}$$

13

14 
$$= \left( \hat{\beta}_1 z_{01}^* + \sum_{j=2}^{D-1} \hat{\beta}_j z_{0j} \right) - \left( \hat{\beta}_1 z_{01} + \sum_{j=2}^{D-1} \hat{\beta}_j z_{0j} \right)$$

15

16 
$$= \hat{\beta}_1 (z_{01}^* - z_{01})$$

17

18 
$$= \hat{\beta}_1 \left( \sqrt{\frac{D-1}{D}} \ln \left( \frac{(1+r)}{(1-s)} \frac{x_{01}}{(\prod_{k=2}^D x_{0k})^{1/(D-1)}} \right) - \sqrt{\frac{D-1}{D}} \ln \left( \frac{x_{01}}{(\prod_{k=2}^D x_{0k})^{1/(D-1)}} \right) \right)$$

1

$$2 \quad = \hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left( \frac{(1+r) x_{01}}{(1-s) (\prod_{k=2}^D x_{0k})^{1/(D-1)}} / \frac{x_{01}}{(\prod_{k=2}^D x_{0k})^{1/(D-1)}} \right)$$

3

$$4 \quad = \hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \text{ as shown in equation (8) of the main paper.}$$

5

6

7

8

### 9 **Confidence interval for $\Delta \hat{y}$**

10

11 The standard error of  $\Delta \hat{y}$  is

12

$$13 \quad SE(\Delta \hat{y}) = \sigma \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \sqrt{(Z^T Z)^{-1}_{(1)}}$$

14 as

15

$$16 \quad \text{var}(\Delta \hat{y}) = \text{var} \left( \hat{\beta}_1 \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \right)$$

1

$$= \left( \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \right)^2 \text{var} (\hat{\beta}_1)$$

2

3

$$= \sigma^2 \left( \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \right)^2 (Z^T Z)_{(1)}^{-1}$$

4

5 where  $Z$  is the design matrix of the model in (1) and  $A_{(1)}$  denotes the first diagonal  
6 element of  $A$ .

7

8 Therefore the  $100(1 - \alpha)\%$  confidence interval for  $\Delta y$  is

9

10

$$\left( \Delta \hat{y} - t_{\alpha/2, n-D} \sigma \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \sqrt{(Z^T Z)_{(1)}^{-1}}, \Delta \hat{y} + t_{\alpha/2, n-D} \sigma \sqrt{\frac{D-1}{D}} \ln \left( \frac{1+r}{1-s} \right) \sqrt{(Z^T Z)_{(1)}^{-1}} \right).$$

11

1 **Supplementary file 2**

2 **Daily activity composition expressed as isometric log-ratio coordinates**

3 A four-part daily activity composition consisting of: sleep (*sleep*), sedentary time  
4 (*SED*), light-intensity physical activity (*LPA*) and moderate-to-vigorous-intensity  
5 physical activity (*MVPA*), can be expressed by a set of three isometric log-ratio  
6 coordinates [  $z_{i1}, z_{i2}, z_{i3}$  ] as follows:

7

8 
$$z_{i1} = \sqrt{\frac{3}{4}} \ln \left( \frac{sleep_i}{\sqrt[3]{SED_i \cdot LPA_i \cdot MVPA_i}} \right),$$

9 
$$z_{i2} = \sqrt{\frac{2}{3}} \ln \left( \frac{SED_i}{\sqrt{LPA_i \cdot MVPA_i}} \right) \text{ and}$$

10 
$$z_{i3} = \sqrt{\frac{1}{2}} \ln \left( \frac{LPA_i}{MVPA_i} \right).$$

11

12

13 To estimate an outcome (e.g., body mass index z-score [zBMI]), a multiple linear  
14 regression model can be constructed with the above log-ratio coordinates as the  
15 explanatory variables:

16

17 
$$zBMI_i = \beta_0 + \beta_1 z_{i1} + \beta_2 z_{i2} + \beta_3 z_{i3} + covariates_i + \epsilon_i$$

18



1 The coefficient  $\beta_1$  corresponds to  $z_{i1}$ , which is the log-ratio of  $sleep_i$ , to the geometric  
2 mean of the remaining behaviours ( $SED_i$ ,  $LPA_i$  and  $MVPA_i$ ).

3

4 Permutation of the compositional parts iteratively to place each behaviour as the first  
5 part of the composition, and then applying the above isometric log-ratio coordinates  
6 allows four linear models to be created, with each model's  $z_{i1}$  representing the first  
7 (permuted) behaviour in relation to the geometric mean of the remaining behaviours.

8

9 *Composition 1 = [Sleep, SED, LPA, MVPA]*

10 *Composition 2 = [SED, LPA, MVPA, sleep]*

11 *Composition 3 = [LPA, MVPA, sleep, SED]*

12 *Composition 4 = [MVPA, sleep, SED, LPA]*

13

#### 14 **Predicting change in zBMI using the linear model**

15 The predictive model using Composition 1 becomes:

16 
$$\widehat{zBMI}_i = \hat{\beta}_0 + \hat{\beta}_1 z_{i1} + \hat{\beta}_2 z_{i2} + \hat{\beta}_3 z_{i3} + covariates_i.$$

17

18 We define the *ilr* coordinates  $[ z_{i1}, z_{i2}, z_{i3} ]$  as described above.

19

1 As compositional data are relative, the predictions from the model must be made  
2 relative to a starting/reference activity behaviour composition. In our example, we select  
3 the population mean activity behaviour composition as the starting point, expressed as  
4 proportions with a closure constant of 1.

5  
6 To predict zBMI for a new composition where our first compositional part (*sleep*) has  
7 changed, we multiply  $sleep_{mean}$  by a constant  $(1 +/- r)$  (e.g., to increase  $sleep_{mean}$  by 5%,  
8  $r = 0.05$ , and we multiply  $sleep_{mean}$  by  $1 + r = 1.05$ ). However, due to the constant sum  
9 constraint of daily activity data, the remaining behaviours must be changed accordingly.  
10 The remaining compositional parts are therefore all simultaneously multiplied by  
11 another constant  $(1 +/- s)$ , specifically derived to maintain the total sum of all parts to 1  
12 (see Supplementary file 1 for details). By multiplying all the remaining parts by the  
13 same constant (in our example the remaining parts are decreased, therefore each  
14 remaining part is multiplied by  $1 - s$ ), the remaining log ratio coordinates  $(z_{i2}, z_{i3})$  are  
15 kept constant (as both numerator and denominator of the log ratio coordinates are  
16 multiplied by the same amount  $[1 - s]$ ).

17  
18 Therefore, we can use the linear model to predict zBMI for a change in the daily activity  
19 composition by the constant  $k$ ,

20 where  $k = \frac{1+r}{1-s}$

1 and  $r$  and  $s$  are defined as described in Supplementary file 1.

2

3 The predictive model for our example becomes (N.B.: the following is a worked

4 example of the proofs given in Supplementary file 1):

5

$$\widehat{zBMI}(k) = \hat{\beta}_0 + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln \left( \frac{sleep}{\sqrt[3]{SED \cdot LPA \cdot MVPA}} \cdot k \right) + \hat{\beta}_2 \sqrt{\frac{2}{3}} \ln \left( \frac{SED}{\sqrt{LPA \cdot MVPA}} \right) + \hat{\beta}_3 \sqrt{\frac{1}{2}} \ln \left( \frac{LPA}{MVPA} \right)$$

7

$$= 8\hat{\beta}_0 + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln \left( \frac{sleep}{\sqrt[3]{SED \cdot LPA \cdot MVPA}} \right) + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln k + \hat{\beta}_2 \sqrt{\frac{2}{3}} \ln \left( \frac{SED}{\sqrt{LPA \cdot MVPA}} \right) + \hat{\beta}_3 \sqrt{\frac{1}{2}} \ln \left( \frac{LPA}{MVPA} \right)$$

9 (*expanding out the first ratio using log law :  $\ln(ab) = \ln(a) + \ln(b)$* )

10

$$\hat{\beta}_0 + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln \left( \frac{sleep}{\sqrt[3]{SED \cdot LPA \cdot MVPA}} \right) + \hat{\beta}_2 \sqrt{\frac{2}{3}} \ln \left( \frac{SED}{\sqrt{LPA \cdot MVPA}} \right) + \hat{\beta}_3 \sqrt{\frac{1}{2}} \ln \left( \frac{LPA}{MVPA} \right) + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln k$$

12 (*rearrange the order so that it's easy to see the first bit can be replaced by 'zBMI(k)'*)

$$= \widehat{zBMI} + \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln k$$

14 Therefore, the predicted change in zBMI is  $\hat{\beta}_1 \sqrt{\frac{3}{4}} \ln k$ .

15

1 Here  $k = \frac{1+r}{1-s}$ , i.e., with an increase in the first behaviour (sleep), *sleep* is multiplied by  
2  $1 + r$ , and the remaining behaviours (*SED, LPA, MVPA*), represented by their geometric  
3 mean, are each multiplied by  $1 - s$ .

4

5 As stated earlier, by multiplying each of the remaining behaviours simultaneously by  $1 -$   
6  $s$ , it is assumed that the log-ratio coefficients  $z_{i2}$  and  $z_{i3}$  are held constant.

7

8 The value of  $1 - s$  can be derived from  $1 + r$ , when we consider daily activity data to be  
9 constrained to  $C = 1$ , i.e., the daily activity components are expressed as proportions.

10

11 The reference/starting daily composition can be expressed as:

12

$$13 \quad \textit{sleep} + \textit{remaining} = 1.$$

14

15 However, we would like to predict an outcome (zBMI) for a new composition, where  
16 sleep is increased by a factor of  $1+r$ , and remaining behaviours are decreased by a factor  
17 of  $1-s$ . The new composition can be expressed as:

18

$$19 \quad \textit{sleep} (1 + r) + \textit{remaining} (1 - s) = 1;$$

20

$$\therefore \textit{sleep} + r \cdot \textit{sleep} + \textit{remaining} - \textit{remaining} \cdot s = 1$$

1 
$$\therefore \text{sleep} + \text{remaining} + r \cdot \text{sleep} - \text{remaining} \cdot s = 1.$$

2

3 As  $\text{sleep} + \text{remaining} = 1$ , therefore:

4

5 
$$(1 + r) \cdot \text{sleep} - \text{remaining} \cdot s = 1;$$

6 
$$\therefore r \cdot \text{sleep} - \text{remaining} \cdot s = 0$$

7 
$$\therefore \text{remaining} \cdot s = r \cdot \text{sleep};$$

8 
$$\therefore s = r \cdot \frac{\text{sleep}}{\text{remaining}};$$

9

10 **Application to the data presented in the manuscript:**

11 The mean daily activity composition (described by geometric means, closed to 1440  
12 minutes) was:

13 
$$\text{sleep} = 539; \text{SED} = 525; \text{LPA} = 320; \text{MVPA} = 57.$$

14

15 This can be expressed as a set of proportions [0.374, 0.364, 0.222, 0.040], which are  
16 closed to 1.

17

18 Expressed as a proportion,  $\text{sleep}_{\text{mean}} = 0.374$ , therefore the *remaining* components  
19 (expressed as proportions) must together equal: 1 (*the total day*) - 0.374 ( $\text{sleep}_{\text{mean}}$ ).

20

1 If we are interested in change in zBMI when sleep is increased (relatively) by 5% ( $r =$   
 2 0.05), we can calculate  $s$ , using the formula above. Specifically,  $s = 0.05 \cdot \frac{0.374}{1-0.374} =$   
 3 0.03.

4

5 We now have our constant  $r (=0.05)$  which is the relative increase in  $sleep_{mean}$ , and our  
 6 constant  $s (=0.03)$ , which is the relative decrease in each of the remaining behaviours.

7 Using  $r$  and  $s$ , we can create  $k$ , the constant which is applied to the first log ratio

8 coordinate ( $z_{i1}$ ). In our example,  $k = \frac{1+r}{1-s} = \frac{1.05}{0.97}$ .

9

10 Earlier we showed that predicted change in zBMI was equal to:

11 
$$\Delta z\widehat{BMI} = \hat{\beta}_1 \cdot \sqrt{\frac{3}{4}} \cdot \ln k, \text{ where } k = \frac{1+r}{1-s}.$$

12 Now, we substitute  $k = \frac{1.05}{0.97}$  into our “change” equation, and use the value for  $\hat{\beta}_1 =$

13  $-0.82$  from the  $ilr_{sleep}$  linear regression model fit (see Table 1 in main paper).

14

15 The change in  $z\widehat{BMI}$  is

16 
$$\hat{\beta}_1 \sqrt{\frac{3}{4}} \ln \frac{1.05}{0.97} = -0.82 \cdot \sqrt{\frac{3}{4}} \cdot \ln \frac{1.05}{0.97} = -0.056.$$

17

18

1 Therefore, predicted zBMI has decreased by 0.056 units when sleep is increased from  
2 the reference/starting composition (in this case, the population-mean composition) by  
3 1.05 or 5%, relative to remaining behaviours.

4

5 The change in daily activity composition can be interpreted in minutes. Mean *sleep* =  
6 539 minutes, therefore a 5% relative decrease in sleep is a decrease of 27 minutes.

7

8 Conversely, if sleep is decreased in relative terms by 5% ( $r = -0.05$ ), then

9

10 
$$k = \frac{1-0.05}{1+0.03}$$

11

12 Therefore,

13 
$$\text{change in } \widehat{zBMI} = -0.82 \cdot \sqrt{\frac{3}{4}} \cdot \ln \frac{1-0.05}{1+0.03} = 0.057.$$

14

15 zBMI is predicted to increase by 0.057 when sleep is decreased by 5% from the  
16 reference/starting composition, relative to remaining behaviours.

17

18

19

1 We can use the linear model to predict how much the first compositional part must  
2 change for a specific change in a continuous outcome (e.g., using the composition  
3 above, we can predict the change in *sleep<sub>mean</sub>* associated with a decrease in zBMI of 0.1  
4 units). To use the model for this prediction, we must first isolate  $r$ .

5

6 We have established that:

7 
$$\Delta \hat{y} \text{ (e.g. zBMI)} = \hat{\beta}_1 \sqrt{\frac{3}{4}} \ln k,$$

8 where (for a 4-part composition),  $y$  = the predicted variable (e.g., zBMI),  $\hat{\beta}_1$  = the  
9 coefficient of the first log-ratio regression coefficient (one behaviour: remaining day,  
10 therefore contains all relative information regarding  $\bar{x}_1$ , the first compositional part of  
11 the mean composition), and  $k = (1+r)/(1-s)$ , where  $s = r(\bar{x}_1x/1-\bar{x}_1x)$ .

12

13 Therefore, rearranging this formula, we can isolate  $k$

14

15 
$$\ln k = \frac{\Delta \hat{y}}{\hat{\beta}_1 \cdot \sqrt{\frac{3}{4}}}$$

16

17 
$$\therefore k = e^{\frac{\Delta \hat{y}}{\hat{\beta}_1 \cdot \sqrt{\frac{3}{4}}}}$$

18



1 To calculate  $r$ , we can use that  $k = (1+r)/(1-s)$ , where  $s = r\bar{x}_1/(1 - \bar{x}_1)$ .

2

3 Therefore,

4 
$$k = \frac{1 + r}{1 - r \left( \frac{\bar{x}_1}{1 - \bar{x}_1} \right)}$$

5

6 where we can isolate  $r$ :

7 
$$r = \frac{-1 + k}{k \left( \frac{\bar{x}_1}{1 - \bar{x}_1} \right) + 1}.$$

8

9 Now we can calculate  $r$  by substituting in  $k$  from above, using the  $\hat{\beta}_1$  from the  
10 regression model, and the  $\Delta\hat{y}$  of interest to calculate  $k$ .

11

12 To express  $r$  as change in minutes from a reference  $x_1$ , e.g.,  $\bar{x}_1$ ,  $r(\text{minutes}) = r \cdot \bar{x}_1 \cdot$

13 1440.

14