ANDS Project

Transformation of the RO dataset: Additional Data Cleaning and Augmentation Procedures

De-identification
Dates of birth were converted to age. Please note: individuals outside the target age range (16-25, *N*=3) were not removed from the dataset.

GPS Data
Several apps used by participants within the Young and Well Towns Cohort Study collected information on the participants location through GPS sensors. Thus longitude and latitude coordinates were collected in regular intervals. However, due to the sensitive nature of this information, this data had to be anonymised to be included in this collection. The following procedure was used to anonymise GPS Data, location coordinates are not provided in the dataset:

For individuals with multiple location data points, the distance to the previous location in m was calculated for each location. This information may be used e.g. to derive information on distances of travel during time intervals or speed of travel. Distance was calculated using the Haversine equation. A mean earth's radius *R* was assumed with *R* = 6371000m. Elevation was not recorded and is not considered in this equation.

Haversine Formula (from R.W. Sinnott, "Virtues of the Haversine", Sky and Telescope, vol. 68, no. 2, 1984, p. 159):

$a = \sin^2(\Delta\varphi/2) + \cos\varphi1 \cdot \cos\varphi2 \cdot \sin^2(\Delta\lambda/2)$
$c = 2 \cdot \text{atan2}(\sqrt{a}, \sqrt{(1-a)})$
$d = R \cdot c$

In addition, for each coordinate, information on places of of interest in the immediate vicinity of the location were collected from an online map service. Data on places of interest (POIs) is provided for POIs within a radius of 250m within the recorded location of the individual. Data is provided as two variables. The first variable *places_count* (integer) represents the number of POIs identified in the vicinity of the location. The second variable *places_data* (string) includes detailed information on the number of specific POIs in JSON format. The number of POIs within this data adds up to *places_count*. Example:

places_count:
6
places_data:
{"types":{"colloquial_area":1,"locality":2,"political":2,"store":1,"point_of_interest":4,"establishment":4,"health":2}}

Removal of 'Test' Data
All 'test' user accounts were removed from the datasets. These can be described as accounts which researchers on the project used to test functioning of the study platform and study measures before and during the RO study. Test accounts were identified and removed based on the sign up name and/or email address. The unique ids for these are listed below:

activity_data:
1057
1059
1063

1055
1188
1021

goAct_monitoring:
570 – 1021
1048
1055
1057
1059
1063
1185
1188

location:
1059
1063
1188
1055

participants:
1056
1057
1059
1063
1188
570-858 (<858)
1021
1044
1045
1046
1048
1049
1053
1054
1055
1184
1185

sleep:
1057
1063
1188
bmi:
1188

body_fat:
1188

location:
1059
1063
1188
1055

RO cohort profiling survey:

R_3njVOPddJm1h36y
R_10voPgskoNYtnjW
R_DItZzZgtl5Ab8tz

BFI-10 data capturing issue
Due to a technical error, one item of the BFI-10 was not captured. This affects the item “I see myself as someone who has few artistic interests”, i.e. one of the items capturing openness.