Exercise 1 Day 3 Advanced Survey Design Summer Course, Utrecht University
Bella Struminskaya & Peter Lugtig

Selection bias in (designed) big data

In the lecture, we talked about the paper by Althoff et al. (2017) published in Nature. The researchers used the data from a smartphone app (Azumio Argus app for iPhones) that tracks people's steps to compare the levels of physical activity across countries and study inequality in physical activity.

In groups of 3-4 people, your task for this exercise is to discuss potential biases that can arise due to selectivity. Take a look at the paper, both the main text and the Methods section (starting on page 5). By selectivity we mean errors of nonobservation on the representation side of the Total Survey Error such as coverage, sampling, nonresponse, and adjustment.

(1) Discuss what biases can arise due to selectivity, you can focus both on:
   - Selectivity occurring prior to being included in the analytic sample
   - Further potential selectivity due to being excluded because of not meeting the criteria set by the researchers.

(2) Propose a designed big data approach: whether and how could researchers have combined surveys with physical activity measurement? What are potential selectivity biases in this approach? How, if at all, is this approach better than the approach of Althoff et al.?

Bonus [if you have time]: you can discuss the validity of measurement of physical activity using step count.