Statistical modeling of pooled data across IPEN countries

The main scope of the IPEN project is to examine environment-physical activity relationships by using data from different cities representing IPEN countries. This entails the conduct of pooled analyses across cities. This statistical guide has the purpose of providing a brief account of approaches to data analyses appropriate for the IPEN project.

The IPEN datasets will have a hierarchical structure and consist of observations nested within census units, census units nested within neighborhoods, and neighborhoods nested within cities representing IPEN countries. This type of sampling requires the use of statistical methods that can account for the dependency of data collected within specific geographical areas (cities, neighborhoods and/or census units). Depending on the aim and scope of the paper and available resources (software and analyst's skills), authors should use one of the following approaches:

1. **Generalized linear models (GLM) with robust standard errors**

They allow modeling of data with diverse distributional assumptions (normal and non-normal). Robust standard errors can account for violations of the independency assumption. When using this modeling approach, neighborhoods or census units would be defined as clusters, while cities would be treated as strata or covariates. GLM provide information on population-averaged effects across all areas (neighborhoods and/or census units across cities). Differential effects across cities can be examined using appropriate interaction terms. This type of modeling approach does not allow simultaneous adjustment for multiple levels of dependency (i.e., neighborhood and census unit level), unless software for complex surveys is used (e.g., SUDAAN, SPSS Complex Samples module, SVY commands in Stata, SURVEY procedures in SAS).

2. **Generalized estimating equations (GEE) with or without robust standard errors**

They also allow modeling of data with various distributional assumptions. The advantage of using GEE over GLM is statistical efficiency (smaller standard errors). However, GEE may not perform well when cluster sizes are highly unbalanced. When using this modeling approach, neighborhoods or census units would be defined as clusters, while cities would be treated as strata or covariates. GEE provide information on population-averaged effects across all areas (neighborhoods and/or census units across cities). Differential effects across cities can be identified and examined using appropriate interaction terms. This modeling approach usually does not allow simultaneous adjustment for multiple levels of dependency (i.e., neighborhoods and census units), unless appropriate complex survey software is used (e.g., SUDAAN).

3. **Multilevel models (mixed effects models)**

Depending on the software used, these models can be applied to normally as well as non-normally distributed data. When using multilevel models, neighborhoods and/or census units would be defined as clusters, while cities may be treated as covariates or clusters. This type of modeling can simultaneously account for multiple levels of

dependency in the data, estimate variances at multiple levels (e.g., how much does the effect of access to recreational facilities on physical activity vary across neighborhoods and cities?) and identify correlates of differences in area-level effects (e.g., what explains neighborhood- and city-level differences in effects of access to recreational facilities on physical activity?). Unlike GEE and GLM (see above), multilevel models provide estimates of area-specific effects and how these vary across areas (cities, neighborhoods or census units). This means that the regression coefficients of multilevel models represent the effect of a variable on the outcome in the 'average' city or neighborhood/census units. These regression coefficients are allowed to vary across cities or neighborhood/census unit. In contrast, GEE and GLM regression coefficients represent the fixed average effect of a variable on the outcome across all areas. Unlike GEE, multilevel models perform well even when clusters are highly unbalanced.

Three modeling issues that are relevant to all three statistical approaches mentioned above.

1. Models examining accelerometry data need to include estimates of their reliability as a covariate. These estimates are represented by the reliability coefficient (ICC) that would be observed for a given period of wearing and monitoring time (i.e., number of hours per day and number of days per week, respectively). To compute these ICCs, each participant's wearing time would be examined to calculate the individual's total days and average hours per day. Then, each participant would be assigned an ICC value pre-calculated from published information for every possible combination of days and average hours per day of wearing time. Alternatively, days of monitoring and wearing time may be included in the regression models as covariates. The downside of this approach is lower statistical efficiency due to the inclusion of two rather than one accelerometer-reliability covariates. The reason for including reliability estimates in models of accelerometry data is that they are indicators of the quality of data collection, which, in turn, can affect the magnitude of the observed associations.

2. The selection of neighborhoods/census units within each study site was not based on random sampling. Therefore, models should be adjusted for the variables used to select the neighborhoods/census units (e.g., walkability index and median household income) if these variables are significantly related with the target outcome. Alternatively, sampling weights at the neighborhood or census block level can be incorporated in the analyses.

3. It is important to ascertain if within-city (or/and within-neighborhood) effects of an explanatory variable differ from between-city (or/and between-neighborhood) effects. If this is the case, both effects should be included in the models.