

INTEGRATED AND MARGINALIZED ROMA IN SLOVAKIA

(study background and data description)

NMST551 – Statistický projektový seminář | Winter term 2026/2027

Introduction and motivation

The data come from the PURE-MD study, a nationwide community-based screening project concerned with movement disorders among Roma people in Slovakia. Roma are largely under-represented in movement-disorder research, even though living conditions, access to health care, occupational exposures and chronic psychosocial stress may all contribute to higher neurological risk. The data combine neurological screening with information about various factors (demographic, socioeconomic, health, family, environmental, behavioural, etc.). Participants entered either as patients/persons with a health complaint or through systematic community screening.

There are 707 participants in total: 229 individuals in the file labelled as *integratedData* and 478 subjects in the file labelled as *marginalizedData*. The data were collected in all regions of Slovakia, but the geographical representation is uneven and is concentrated mainly in eastern Slovakia.

Scientific questions

The principal objective is to quantify and explain differences in movement-related symptoms and diagnosed movement-disorder phenotypes between the two study populations—the integrated Roma and the marginalized Roma.

1. *What is the prevalence of various movement disorders in the two Roma subpopulations?*
2. *How are movement problems associated with socioeconomic conditions, housing and infrastructure, environmental or occupational exposures, stressful life events, health-care use, lifestyle and comorbidities?*
3. *Do the clinically assigned categories in EXPY (parkinsonism, ataxia, hyperkinetic disorder or healthy control) agree with the symptom-screening outcomes? Among hyperkinetic cases (determined by EXPY = *hyperkinetické*), what is the distribution of the more specific phenotypes provided in EXPY-H?*

The primary comparison of the integrated and marginalized Roma may use the subpopulations defined by the corresponding Excel files. Alternatively, a marginalized Roma person can be defined as an individual who satisfies at least one of the following: living in a shack ($SE-B = 4$); relying on a stream/creek for drinking water ($M-W = 4$); cooking on a wood stove ($M-C = 1$); having no flushing toilet ($M-T = \text{“nie”}$); or having no functional bathroom ($M-B = \text{“nie”}$). In addition, other well-justified approaches may also be used.

Data files and variables

Two Excel files are provided. The file `integratedData.xlsx` contains 229 participants, while the second file, `marginalizedData.xlsx`, contains 478 individuals. Both files have the same structure. In the first worksheet, participants are stored in *columns* and 125 variables are provided in *rows*. A detailed data description is provided in the second worksheet together with specific groups of categories (social, health, physical, etc.).

Technical and interpretive limitations

Missingness is extensive and strongly variable-specific. Some apparent blanks are structural (for example, follow-up items after “none”), while others represent unknown values, so they must not all be imputed identically. Coding is also heterogeneous: yes/no answers are in Slovak, categorical variables are stored as numerical values, decimal commas occur in some continuous variables, and there are occasional spaces or spelling variants.