

# Stroke patients' rehabilitation in Estonia 2017-2020

GitHub repository: <https://github.com/K-AMeus/IDSPProject>

Team members:

- 1) Kevin Kits
- 2) Richard Prost
- 3) Karl-Andreas Meus

## Task 2. Business understanding

### 1. Identifying Your Business Goals

- 1) Background
  - Stroke is a major public health concern and its impact on individuals and the healthcare system is substantial. In Estonia, between 2017 and 2020 stroke patients underwent rehabilitation processes and understanding and optimizing rehabilitation is vital for patients health and the efficiency of the healthcare system.
- 2) Business goals
  - Find out if it is useful for the state and Põlva hospital to keep stroke rehabilitation in the Estonian healthcare system.
  - Highlighting regional differences and per capita comparisons.
- 3) Business success criteria
  - Põlva hospital can decide whether it is useful for the state to keep stroke rehabilitation in the Estonian healthcare system.

### 2. Assessing the situation

- Inventory of resources:  
  
Stroke patients rehabilitation data,  
3 BSc students,  
1 medical expert,

3 university provided laptops.

- Requirements, assumptions and constraints:

Not exactly clear what data privacy problems we could face (not clear right now) but obviously we have to play it safe.

Data should be complete and we might not get access to further/additional data in time as getting this data was quite a process in itself.

- Risks and contingencies

There could be a social engineering risk factor for our project. As the data is obviously not publicly available and could be sensitive, we have to keep our guard up. Don't know if we can have a backup plan in case that happens.

- Terminology

Stroke types:

Hemorrhagic - Due to bleeding into the brain

Ischemic - Due to cut off blood supply to the brain

- Costs and benefits

Costs: 90h of brainpower

Benefits: Improved healthcare efficiency, predictive analytics for stroke patients' rehabilitation, personalized medicine/treatment for patients (Ideally)

### 3. Data-mining goals

#### 1) Data-mining goals.

- Analyze the differences between active rehabilitation and post-rehabilitation and how they impact patient outcomes.
- Evaluate whether rehabilitation provides benefits to the Estonian healthcare system.
- Display the receipt of therapies over time and compare different types of rehabilitation.
- Highlight regional differences in rehabilitation and conduct a per capita analysis.
- Create a comprehensive visual representation across years and counties to encompass all significant findings.
- Develop a strategy to handle the presence of numerous zeros in the dataset to ensure that it does not lead to misleading information, especially in scenarios where the median might be zero.

## 2) Data-mining success criteria

- Clear identification of differences in outcomes between active and post-rehabilitation.
- Benefits could be associated with reduced post-rehabilitation needs or lower healthcare costs. Use Kaplan-Meier curves for survival analysis.
- Significant differences in the timing of therapy receipt should be noticeable. Compare the impact of different types of therapies on patient outcomes.
- Clearly demonstrate differences in the frequency and quality of rehabilitation across counties.
- Monitor and handle null values' impact on the analysis.
- The graphs should be clear and easily interpretable.
- Highlight key trends and differences across years and counties.

## Task 3. Data understanding

### • Gathering data

#### a. Outline data requirements

We are probably going to need all of the data we can get our hands on. The most important attributes might be: stroke\_type, treatment\_method, age, therapy\_phase. It's possible we find that some other attribute is much more important.

#### b. Verify data availability

The data that we have is not publicly available. This is private data that we got from our medical expert source. As far as we are aware we will get enough data (at this moment we don't have all of the promised data, only a sample).

#### c. Define selection criteria

To narrow down the dataset for our analysis, we will focus on relevant years (2017-2020). Additionally, we may filter out records with missing or inconsistent data, ensuring the reliability of our findings (null attributes).

### • Describing data.

The dataset consists of stroke patient rehabilitation records from 2017 to 2020 in Estonia. In the dataset we have attributes: Id (patient identifier), age, sex, county, comorbidity (the simultaneous presence of two or more diseases or medical conditions in a patient), stroke\_type (ischemic or hemorrhagic), dementia, year, treatment\_method, therapy\_phase\_I (active treatment), therapy\_phase\_II (follow-up treatment), time (time of death), status (alive or dead).

- **Exploring data**

We will explore our data deeper when we receive it fully. From the current data example we have we can see the attribute data types.

- **Id** - integer
- **Age** - integer
- **Sex** - char 'f' or 'm'
- **County** - string
- **Comorbidity** - integer
- **Stroke\_type** - string (ischemic or hemorrhagic)
- **Dementia** - string (yes or no)
- **Year** - integer
- **Treatment\_method** - string (surgical or nonsurgical)
- **therapy\_phase\_I** - float
- **therapy\_phase\_II** - float
- **Time** - float
- **Status** - integer (0 or 1)

- **Verifying data quality**

We will verify our data once we receive it fully.

## **Task 4. Planning your project**

### **1. Data Collection (3h/student)**

- a. Figure out what we want to collect
- b. Find a source for data
- c. Get permission to use data
- d. Get data

### **2. Data Cleaning (5h/student)**

- a. Figure out our preprocessing.
- b. Clean the data
- c. Handle missing values

### **3. Exploratory Data Analysis (5h/student)**

- a. Generate descriptive statistics and visualizations for key attributes.
- b. Conduct correlation analysis between variables.
- c. Find promising patterns
- d. Find anomalies and try to interpret

#### **4. Model Development (5h/student)**

- a. Choose appropriate models for predicting rehabilitation outcomes.

#### **5. Model Evaluation and Validation (6h/student)**

- a. Assess model performance using metrics like accuracy, precision, and recall.

#### **6. Documentation and Reporting (6h/student)**

- a. Summarize findings, insights, and recommendations.
- b. Prepare a comprehensive report outlining the entire process.

#### **TOOLS:**

1) Python (Pandas, Numpy, Matplotlib, Seaborn)

2) Jupyter notebook

3) Github, Trello