# Forecasting the physical fitness and all-cause mortality based of schoolchildren's fitness measurements

Matej Cigale
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenija
matej.cigale@ijs.si

Anton Gradišek
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenija
anton.gradisek@ijs.si

Miha Mlakar
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenija
miha.mlakar@ijs.si

Mitja Luštrek
Jožef Stefan Institute
Jamova cesta 39
Ljubljana, Slovenija
mitja.lustrek@ijs.si

## ABSTRACT

The focus of medicine is steadily shifting from curing the sick to preventive measures. In order to assist the policy makers in making the right decisions that would lead to a healthier population, there is an increasing need to develop models that can forecast the state of the population in the future, check what measures are effective and what policies synchronize. In order to track these changes, predict the state of the population in the future, and thus make informed decisions, the CrowdHEALTH platform is developed. The SLOfit use case takes the information collected from a large population of school children and aggregates this to provide information on the future health of the population that is just now finishing school.

## Keywords

obesity, fitness, exercise, machine learning

## 1. INTRODUCTION

The focus of medicine is steadily shifting from curing the sick to preventive measures[8]. With people's growing desire to increase their lifespan and health, there is an ever greater push for the policy makers to provide ways for people to increase or maintain their fitness. In order to forecast what the population health will be in the years to come, research looks to the machine learning algorithms that can generate models predicting the trends in specific populations. The focus of the CrowdHEALTH (CH) project is to gather this kind of information in a consistent way across multiple data sources and generate models that can be used to predict what the effects of implementing health policies will be on a population.

SLOFit is a large study on physical fitness that includes data collected from Slovenian schoolchildren for over 40 years, and is used to chart global health trends in the population. This data set was used as the basis for our modeling so that we could predict the state of the population in the future (when they are grown) and calculate what the associated risks for mortality will be in the future. In the course of the project we investigated several models that can be used to predict the state of the population and, as expected, the prediction of fitness for children based on previous years is quite hard. Several approaches were evaluated, but at this time linear regression seems to provide the best results, although research on creating better models is still ongoing. Since data on the health risks for the subjects in the SLOFit dataset is not available, we use risk calculation based on the literature to calculate general mortality models based on certain fitness indicators. As the data stored in the system and its applications is complex, the standard used must enable flexible storage of information. The CH infrastructure uses Fast Healthcare Interoperability Resources (FHIR) [3] as the standard for data storage, meaning that all data can be queried in a similar manner, and if the appropriate information is available, compared and forecasts generated.

The rest of the paper is organized as follows. Section 2 provides an overview of the SLOFit data set. Section 3 provides the information on the architecture of the Forecasting Analytical tool and places it in the context of the CH system. Section 4 discusses the forecasting algorithm. Section 5 provides the outline of the Risk assessment and finally Section 6 discusses the results.

## 2. SLOFIT DATASET

SLOfit is a massive cohort study of physical fitness of Slovenian schoolchildren. Every April, almost all elementary and high school students undergo measurements of 3 anthropometric tests (height, weight, triceps skinfold) and 8 motoric tests, aimed at monitoring different components of physical fitness (such as cardiorespiratory fitness, muscular fitness, explosivity, agility, coordination, etc.) The SLOfit study has been ongoing on the national level since 1987 and serves as the scientific backbone for most of policies related to physical education in schools and enhancing of physical fitness in schoolchildren. To date, the SLOfit database includes over 7 million sets of measurements for over 1 million children, being one of the largest cross-sectional and cohort databases of physical and motor development in the world.

In our study on forecasting of physical fitness, a subset of the SLOfit data was used, encompassing the data from aproximately 2000 children from the age of 6 to 18. In the analysis,

the data was anonymized, retaining only the municipality-level data in order to be able to create policies on regional level. When assessing risks, we focus on a subset of SLOfit parameters that are directly connected to the risks we are interested in. Height and weight are used to calculate the body-mass-index (BMI), which is used to determine whether a person is overweight (obesity) or underweight. 30 s sit-up results are used as a proxy for muscular fitness (MF), while the 600 m run results are indicative of cardiorespiratory fitness (CRF). In the risk analysis, we are currently focusing on all-case early mortality risks while risks for developing cardiovascular diseases (CVD) or diabetes are planned to be looked at in future.

## 3. ARCHITECTURE OVERVIEW

The data in the CH project is stored following an extension of the FHIR standard, where each measurement is stored as an observation that includes all the meta-data of the measurement, such as when it was taken, by whom, what are the units of the measurement, etc. The current architecture of the CH system is demonstrated in Figure 1. This enables the overall system to be extended in the future with custom tools.

The data is stored in LeanXcale (LXC) [1], a flexible, ultra-scalable database with analytical capabilities. In order that the information of different types can be stored, a specialized schema was developed. The part that is pertinent to our work was the addition of a new Person class - Student, to differentiate it from Patient that is the general subject in FHIR. Additionally, the metadata for schools, municipalities and regions were added. The 11 standard anthropometric and motoric tests were also codified in the system so that they can be easily accessed. In order to speed up the queries, the Forecasting module we developed includes a small internal database that caches the data. This is facilitated by SQLite with schema that mimics the data stored in LXC system - i.e., the region, municipality, school, student and observation classes that include most of the data stored in the LXC system. This provides faster look-up times and simplifies some filtering, as SQLite can be tightly coupled with the Django service.

Django is a framework that enables the creation of web APIs in Python. It consists of three main parts. Django Models are mapped directly to supported databases, allowing for fast and efficient filtering and querying of the system. The developer can specify the DB schema and provide rules to check if the values are correct, serialize the data and link the tables in several ways. Djanogo Models are specification that maps directly to the DB schema, that the Framework actually creates for itself, and also handles creation of queries to the system. The Django Views are where the processing of the data happens. Each request can be handled here and responded to accordingly. Django Templates are the presentation layer of the system, but are not used in the current implementation as this is handled by outside systems.

## 4. FORECASTING

The task of the forecasting algorithm is to predict a particular SLOfit parameter (height, weight, sit-ups, 600 m run) at the age of 18, based on the data from previous years and knowing the general population trends.

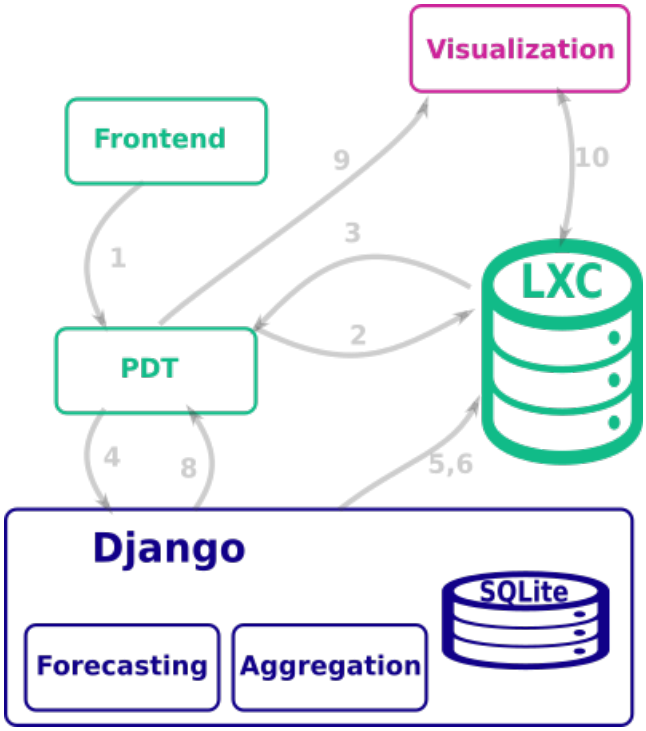


**Figure 1: The architecture and flow of the application.**

As the starting point, we defined two baseline forecasting approaches. The first one, called baseline percentiles, uses the percentile method: if an individual is in the $n$-th percentile at the age of 13, we assume he would be in the same percentile at the age of 18. The second baseline model, called baseline average growth, uses the current value and adds the average growth values for each year until the age of 18.

More advanced approaches use machine learning. To improve the prediction accuracy, we generated additional features, such as average, maximum and minimum year growth, standard deviations, data percentiles and peak height velocity - PHV (the year with the maximum growth). Since PHV was not notated in data, we had to estimate it. We manually annotated it for the small amount of children and then trained a prediction model for PHV on this data. We used this model to predict PHV on all other data. These predictions are not 100 % accurate, but this information as an input for predicting e.g. height improves predictions as will be seen in the results.

Next, we built a model for each year up to which we have available data. For example, the model for the age of 13 takes the measurements from ages 6 to 13 and forecasts the value at the age of 18. Since we have data from 6 to 18 years, we build 12 models for each SLOfit parameter.

Additionally, we enriched each SLOfit parameter data with additional data from another parameter. For predicting the height, we also used weight, for predicting weight, we also used height, for predicting sit-ups we also used results from the 600 m run, and when predicting the 600 m run, we also

used data from the 60 m run.

Several machine-learning algorithms were tested on the data set of 2000 children introduced in Section 2. To evaluate and compared them, the average absolute error was calculated for each years' predictions and then the average error over all the years. This average error over all the years for predicting the height is presented in Table 1.

**Table 1: Comparing algorithms when predicting height.**

| Method | Average error [mm] |
|---|---|
| Baseline percentiles | 36.0 |
| Baseline average growth | 34.3 |
| Linear regression | 27.5 |
| Decision tree | 38.9 |
| Logistic regression | 41.2 |
| SVM | 52.3 |

As we see, the best results were obtained using a linear regression model. Very similar results were obtained also when predicting other SLOfit parameters. The average errors for each year for linear regression and baseline models are shown in Figures 2–5. We see that predicting the values at the age of 18 is a hard problem, so the errors start to decline just a few years before this age.

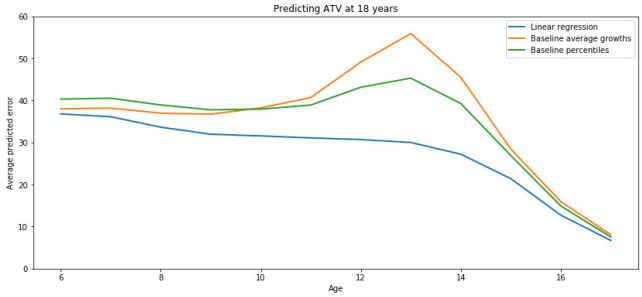


**Figure 2: Prediction error for each year when predicting height at the age of 18.**

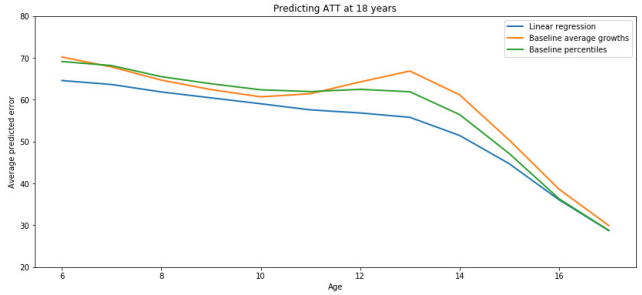


**Figure 3: Prediction error for each year when predicting weight at the age of 18.**

## 4.1 Forecasting the population based on collected measurements

Forecasting of SLOfit parameters takes place at the level of an individual. The available measurements of a person are

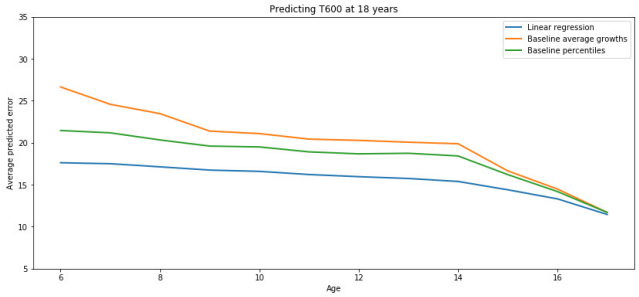


**Figure 4: Prediction error for each year when predicting time running 600 meters at the age of 18.**

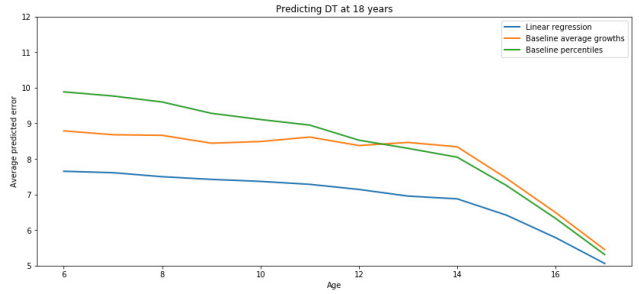


**Figure 5: Prediction error for each year when predicting sit-ups at the age of 18.**

taken from the database and using an appropriate model, predictions at the age of 18 – we assume the person will be fully developed by this age – are generated. Models are generated for each SLOfit parameter and as such need to be run separately for each prediction we want to generate.

As the overall goal of the system is to generate the predictions for a group of people, not just for an individual, the result should be a cross-section of the population based on a filter that is applied. The filter is usually the mean for the population, but other options are also available, for instance quartiles or median. The system automatically takes the information for the available children based on the region where they are from, and generates forecasts for each child. This can require the system to generate multiple forecasts for each individual, for instance height and weight if the desired outcome of the analysis is the BMI. Due to the nature of the system the result of this operation is stored in the database and must be retrieved from there. The aggregator then takes this information and generates reports that can be visualized by the CH systems.

# 5. RISK ASSESSMENT

In order to asses the risks for mortality, a stochastic model was generated that describes how BMI, CRF (approximated by 600 m run) and MF (approximated by sit-ups in 30 s) influence mortality. The influence is based on several published studies [2, 4, 5, 7] that relate fitness indicators to all-cause mortality. Table 4 shows how less-than-ideal values of different parameters increase the probability of mortality [6]. In the case of BMI, it is not surprising that this happens if the individual is overweight. But low BMI is also a risk as it signals other difficulties of the person. The risk for obese

**Table 2: The risk increases for certain calculated metrics.**

| | | | | | |
|---|---|---|---|---|---|
| BMI (kg/m2) | 15-18.5 | 18.5-20 | 20-22.5 | 22.5-25 | 25-27.5 |
| Risk increase(%) | 82 | 44 | 2 | ref. | 7 |
| BMI (kg/m2) | 27.5-30 | 30-35 | 35-40 | 40-60 | |
| Risk increase(%) | 27 | 66 | 166 | 335 | |
| CRF (600m run ) | Q1 (high) | Q2 | Q3 | Q4 | Q5 (low) |
| Risk increase (%) | ref. | 28 | 59 | 78 | 85 |
| MF (30s sit-u ps) | Q1 (high) | Q2 | Q3 | Q4 (low) | |
| Risk increase (%) | ref. | 61 | 32 | 172 | |

people rises quite drastically, since increased weight prevents a person from exercising, further decreasing fitness and increasing the risk for comorbidities of physical or psychological nature. Low CRF and MF have similar consequences, increasing the risk to one's health. While these factors are certainly correlated there is at this time no quantitative data to what extent the correlation should be taken into account. There is also no concrete information how fitness at the end of schooling predicts the fitness of individuals during the rest of their life, as they can at any time decide to change their lifestyle. However, since the change can be for the better or worse, we assume it stays the same, which is probably not far from the truth for the whole population.

# 6. CONCLUSIONS AND FURTHER WORK

Predicting the state of the population and the associated risks for them in the future is an important goal if we want to provide good advice to individuals and people that are directly or indirectly given charge over them. While children are the focus of the current work, the implications are broader. The same approaches could be used on the adult population, predicting their physical fitness and assessing their risks during their lifetime.

In the future work of the project we plan to increase the predictive power of the models by using more data and more advanced machine-learning methods. Risk assessment will be augmented by additional studies from the literature. We would also like to base it on our own data, but it is doubtful we will be able to obtain appropriate data, since most schoolchildren in the SLOfit dataset do not yet suffer from many serious health problems, and relating their fitness with medical data is problematic for privacy reasons.

Perhaps the greater advancement will be achieved by modeling the impact of various health policies and interventions – for instance, what happens if an additional hour of physical education is instituted at a school.

# 7. ACKNOWLEDGMENTS

Funding: This work was supported by the European Union's Horizon 2020 research and innovation program [grant agreement No 727560 (CrowdHEALTH)].

# 8. ADDITIONAL AUTHORS

Additional authors: Maroje Sorić (Faculty of Sports, University of Ljubljana, email: `Maroje.Soric@fsp.uni-lj.si`) and Gregor Starc (Faculty of Sports, University of Ljubljana, email: `Gregor.Starc@fsp.uni-lj.si`) and
Bojan Leskovšek (Faculty of Sports, University of Ljubljana, email: `Bojan.Leskosek@fsp.uni-lj.si`) and Gregor Jurak (Faculty of Sports, University of Ljubljana, email: `Gregor.Jurak@fsp.uni-lj.si`).

# 9. REFERENCES

[1] A. Azqueta-Alzuaz, M. Patino-Martinez, I. Brondino, and R. Jimenez-Peris. Massive data load on distributed database systems over HBase. *Proceedings - 2017 17th IEEE/ACM International Symposium on Cluster, Cloud and Grid Computing, CCGRID 2017*, pages 776–779, 2017.

[2] C. E. Barlow, L. F. DeFina, N. B. Radford, J. D. Berry, K. H. Cooper, W. L. Haskell, L. W. Jones, and S. G. Lakoski. Cardiorespiratory fitness and long-term survival in "low-risk" adults. *Journal of the American Heart Association*, 1(4):e001354, 2012.

[3] D. Bender and K. Sartipi. Hl7 fhir: An agile and restful approach to healthcare information exchange. In *Computer-Based Medical Systems (CBMS), 2013 IEEE 26th International Symposium on*, pages 326–331. IEEE, 2013.

[4] N. R. F. Collaboration et al. Trends in adult body-mass index in 200 countries from 1975 to 2014: a pooled analysis of 1698 population-based measurement studies with 19· 2 million participants. *The Lancet*, 387(10026):1377–1396, 2016.

[5] E. Di Angelantonio, S. N. Bhupathiraju, D. Wormser, P. Gao, S. Kaptoge, A. B. de Gonzalez, B. J. Cairns, R. Huxley, C. L. Jackson, G. Joshy, et al. Body-mass index and all-cause mortality: individual-participant-data meta-analysis of 239 prospective studies in four continents. *The Lancet*, 388(10046):776–786, 2016.

[6] A. Gradišek, M. Mlakar, M. Cigale, L. Lajovic, M. Luštrek, M. Sorić, G. Starc, B. Leskošek, and G. Jurak. Physical Fitness Forecasting and Risk Estimation in Slovenian Schoolchildren. *Studies in health technology and informatics*, 251:125–128, 2018.

[7] P. T. Katzmarzyk and C. L. Craig. Musculoskeletal fitness and risk of mortality. *Medicine and science in sports and exercise*, 34(5):740–744, 2002.

[8] G. Miller, C. Roehrig, P. Hughes-Cromwick, and C. Lake. Quantifying national spending on wellness and prevention. In *Beyond Health Insurance: Public Policy to Improve Health*, pages 1–24. Emerald Group Publishing Limited, 2008.