SmartCHANGE Risk Prediction Tool: Demonstrating Risk Assessment for Children and Youth

Marko Jordan
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
marko.jordan@ijs.si

Nina Reščič
Jožef Stefan Institute,
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
nina.rescic@ijs.si

Sebastjan Kramar
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
sebastjan.kramar@ijs.si

Marcel Založnik
Jožef Stefan Institute,
Department of Intelligent Systems
Ljubljana, Slovenia
marcel.zaloznik@ijs.si

Mitja Luštrek
Jožef Stefan Institute,
Department of Intelligent Systems
Jožef Stefan International
Postgraduate School
Ljubljana, Slovenia
mitja.lustrek@ijs.si

Abstract

Non-communicable diseases (NCDs) have become a significant public health challenge in developed countries, driven by common risk factors such as obesity, low physical activity, and unhealthy lifestyle choices. Early childhood and adolescence are crucial for establishing healthy behaviours, and early intervention can play a crucial role in preventing or delaying the onset of NCDs later in life. However, current tools for identifying high-risk individuals are primarily designed for adults, which results in missed early detection opportunities in younger populations. The SmartCHANGE project (https://smart-change.eu/) seeks to bridge this gap by developing reliable AI tools that assess risk factors in children and adolescents as accurately as possible while promoting optimized risk reduction strategies.

In developing the risk assessment tool, we addressed the challenge of merging diverse datasets, predicting missing data to create longitudinal datasets, implementing existing validated models for diabetes (QxMD) and cardiovascular disease (SCORE2), and ultimately creating a simple online application to demonstrate the functionality of the developed risk tool.

Keywords

risk tool, dataset merge, neural networks, online application

1 Introduction

In developed countries, non-communicable chronic diseases (NCDs) have emerged as the foremost public health challenge over recent decades. According to the World Health Organization (WHO), NCDs account for more than 70% of mortality in the European region [18]. Common risk factors for NCD include obesity, poor physical fitness, and unhealthy lifestyle habits such as inadequate physical activity, sedentary behaviour, poor nutrition, insufficient sleep, smoking, and excessive alcohol consumption. Embracing a healthy lifestyle can improve physical, social, and mental well-being, especially among youth, while mitigating the risks of NCD-related morbidity and mortality [15], [14], [5].

Traditionally, clinical prevention strategies for NCDs have been directed at adults, as the risk factors typically become evident in adulthood. However, recent evidence suggests that focusing interventions on children and adolescents can be a more effective strategy for reducing NCD risk through behaviour modification [13]. While NCDs may not appear in childhood or adolescence, early signs can alreadexistnt. Tackling risk factors and promoting healthy habits during these stages can prevent or delay NCDs later in life [12]. Childhood and youth are also crucial periods for establishing healthy lifestyle habits. Since risk factors for NCDs often persist from childhood into adulthood [9], early risk assessment and reduction of risk factors can potentially lower the incidence of NCD. Lastly, NCDs in youth are a significant global health challenge, with nearly one in five adolescents worldwide being overweight or obese [1].

Identifying high-risk individuals for future health problems is essential for targeted preventive interventions. Existing tools focus mainly on adults [6], for instance predicting 10-year risk of developing cardiovascular disseased [17] or diabetes [8], missing the opportunity to identify high-risk individuals during childhood and adolescence, a critical period for forming lifestyle habits. However, recognition of health risks is not a trivial task. For instance, only 35% of doctors in the UK are aware of the recommendations for physical activity, and only 13% can specify the recommended weekly duration. Moreover, more than 80% of parents of inactive children incorrectly believe that their children are sufficiently active [4]. Developing risk prediction tools for children and youth would significantly improve NCD prevention and promote cost-effective strategies.

This paper presents the development of an initial demo application of a risk assessment tool designed for children and adolescents in the SmartCHANGE project [3] - merging datasets, predicting missing data to build longitudinal datasets, and implementing existing validated models for diabetes (QxMD) and cardiovascular disease (SCORE2) and finally, the application development.

Information Society 2024, 7–11 October 2024, Ljubljana, Slovenia

https://doi.org/10.70314/is.2024.scai.8844

Table 1: Overview of Selected Datasets

Dataset Name	SLOfit	LGS	YFS	AFINA-TE
Country of Origin	SI	BE	FI	PT
Age Range	5 - 20	5 - 25	0 - 60	5 - 25
Longitudinal Study	Yes	Yes	Yes	No
# of Participants	280,165	17,991	3,596	1,632
# of Measurements	3,121,399	31,127	32,364	1,632
# of Variables	13	80	24	59
% of Missing Values	2.55%	16.25%	39.49%	33.53%

2 Methodology

2.1 Datasets

To estimate the risk of non-communicable diseases in children, ideally, one would need a dataset that tracks risk factors from a young age (when the prediction is made) to an older age (when these diseases typically emerge). Such comprehensive longitudinal datasets would allow for accurate predictions of an individual's likelihood of developing a disease later in life based on their early risk factors. However, such datasets are currently unavailable, so we must rely on a collection of partial and often heterogeneous datasets.

In our study, we have chosen 16 types of variables that are used by risk models SCORE2 [17] and QxMD [8]. The datasets we were using are described in Table 1. The SLOfit program is a school fitness monitoring initiative in Slovenia [11]. The Leuven Growth Study (LGS)[2, 16] is a longitudinal study initiated in 1969 that evaluates physical fitness. The Cardiovascular Risk in Young Finns Study (YFS)[10], started in the late 1970s, focuses on early cardiovascular disease risk factors. The AFINA-TE dataset [7] is part of an intervention program in Portugal designed to enhance physical fitness, activity, and nutritional knowledge among children and adolescents.

2.2 Data Imputation Through Datasets

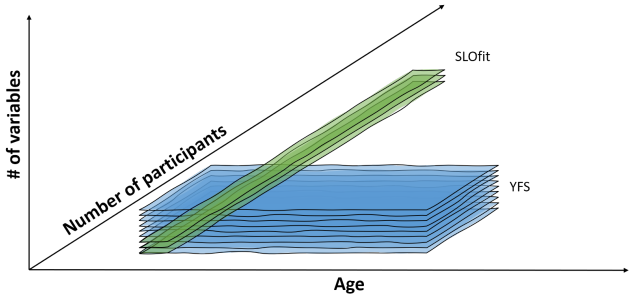
The first step involved imputing missing values within each dataset (see Figure 1 for representation). To guide this process, we calculated the coverage for each variable. Initially, we used only fully observed variables—such as height, weight, and sex—as features in models to impute missing values for other variables. The variables were imputed based on their coverage using machine learning on existing features. After this initial imputation sweep, we had a complete, though potentially imperfect, dataset. In the second sweep, we treated all columns as complete, incorporating the newly imputed values from the first sweep. This allowed us to train models with a more comprehensive dataset, improving the accuracy of the imputation.

2.3 Longitudinal Data Imputation

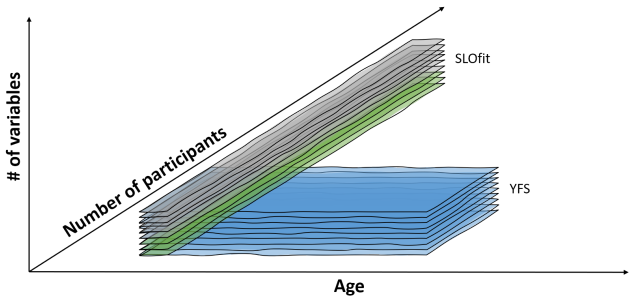
In the second step, we employed a similar approach but focused on merging the datasets to fill the new merged dataset longitudinally (see Figure 2 for representation). To maximize their overlap, we treated certain variables as equivalent—such as vertical jump from the LGS dataset and standing long jump from the SLOfit dataset.

Since the raw values of these variables differ, we standardized them by converting them to z-scores, which were calculated as follows:

$$z\_score = \frac{variable - mean}{standard\_deviation}.$$



(a) Example of the datasets pre-imputation.



(b) Example of the datasets post-imputation.

Figure 1: The YFS dataset (blue) covers a broad range of variables across a wide age span but includes a relatively small number of participants. In contrast, the SLOfit dataset (green) has many participants but includes fewer variables over a shorter age span. In the first step, we imputed the missing variables across the datasets (grey).

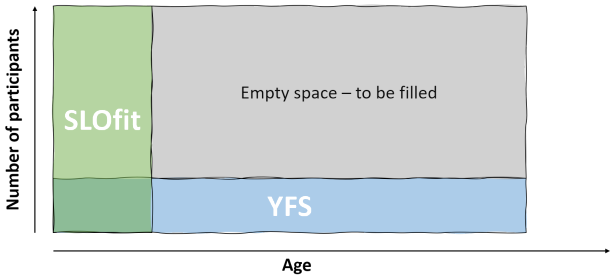


Figure 2: Longitudinal filling of the datasets.

For instance, a vertical jump one standard deviation above the mean in the LGS dataset was considered equivalent to a standing long jump one standard deviation above the mean in the SLOfit dataset. After matching and standardizing the columns across datasets, we merged the individual datasets into a single, comprehensive dataset and repeated the imputation process.

With a merged dataset free of missing values, we built models to predict attribute values at age 55—the oldest age supported by our data—using values from age 14. Due to the lack of data covering the entire age range from 14 to 55, we approached this in two stages: predicting from age 14 to 18 and then from 18 to

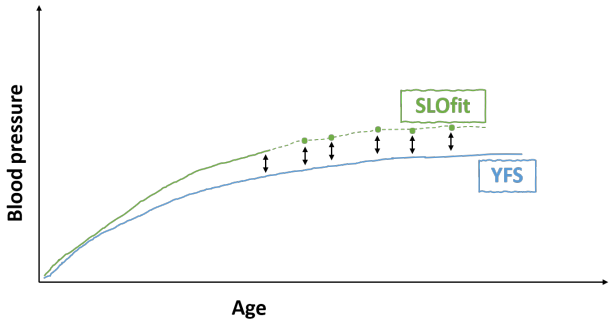


Figure 3: Population-based approach using z-scores.

55. The models used were simple neural networks with a single hidden layer.

This individual forecasting approach required available data for the same person from the start to the end age. However, since we had more data available for different people of various ages, we also explored a population-based approach to forecast the typical evolution of each variable. While this method is less personalized, it is also less prone to anomalies caused by atypical individuals. In the population-based approach, we again used z-scores, assuming that each person's z-score remains constant. For example, if someone's blood pressure is one standard deviation below the mean at age 14, it is assumed to stay one standard deviation below the mean at age 55 (see Figure 3).

2.4 Risk Models

The SCORE2 and QxMD models were used in the application to assess cardiovascular disease and type 2 diabetes risk. These models were chosen for their validity, robustness and effectiveness in predicting these chronic conditions. By incorporating both, healthcare practitioners can comprehensively evaluate cardiometabolic risk factors, aiding in well-informed patient management and intervention decisions.

The SCORE2 model, developed by the European Society of Cardiology, estimates the risk of cardiovascular events over ten years. It calculates the risk score by incorporating variables such as age, sex, smoking status, blood pressure, and lipid profile. Additionally, SCORE2 considers regional variations in risk factors, providing more accurate predictions tailored to specific populations [17].

The QxMD Diabetes Risk Calculator, a comprehensive clinical decision support tool, is employed to evaluate the risk of developing type 2 diabetes mellitus. This model integrates risk factors, including age, BMI, family history, physical activity level, and dietary habits, to estimate an individual's diabetes risk [8].

3 Evaluation

Table 2 presents the cross-validated evaluation results of our forecasting models. As anticipated, the errors in the first stage of individual forecasting are shallow due to the relatively short period. The emistakesin the second stage are higher but still considered acceptable, with the notable exceptions of weight and smoking. We hypothesize that the high variability during puberty, which many adolescents experience around age 14, complicates accurate weight forecasting. In population forecasting, the errors are generally more significant, which aligns with the less personalized nature of this method. However, weight is forecasted

	Ind. 18	Ind. 55	Pop.
Height [cm]	3.11	3.47	1.62
Weight [kg]	4.79	13.60	10.58
SBP [mmHg]	1.46	2.39	10.91
Total cholesterol [mmol/L]	0.05	0.10	0.64
HDL [mmol/L]	0.02	0.08	0.21
LDL [mmol/L]	0.05	0.17	0.51
Smoking [1-9]	1.01	1.72	2.26

Table 2: Mean absolute error for individual forecasting to ages 18 and 55, and for population forecasting.

with greater accuracy in this approach. In the future, we may explore combining both methods or select the more accurate one depending on the variable.

4 Demo Application

To show the general idea of the project, we constructed a demo application implemented with Python in the Dash framework. In the app, a user can specify the inputs (some inputs, such as steroid use, were fixed to make the app more concise) to the models, which in turn yielded two plots which showed how the cardiovascular and diabetes risk evolved from the currently selected age up to an age of an older adult, at age 55. In a different plot, we also showed how a risk factor chosen changes over time.

4.1 Risk Prediction using Demo Application

The developed demo application interface offers a dynamic tool for visualizing health risks based on various user-input parameters used in risk models (Figure 4). By allowing users to adjust these parameters, the dashboard generates real-time projections of two key risk metrics: a 10-year cardiovascular risk score and a 10-year risk of developing diabetes. These risks are shown in two line graphs, illustrating how these conditions' probability evolves with age. Additionally, the dashboard includes a feature that tracks the progression of a selected health parameter (BMI, systolic blood pressure, total cholesterol, HDL) over time, providing insight into how this factor might change as the individual ages. The developed tool intuitively explains how lifestyle and physiological factors contribute to long-term health risks, offering valuable insights for clinical decision-making and personal health management.

4.2 Further Development of the Application

The current version of the demo application is developed based on the data and models currently available. However, there remains an open question regarding the specific needs and preferences of the medical experts who will ultimately use the final application. To address this, we plan to present the current version to these experts and, based on their feedback, refine and enhance the application in subsequent iterations.

5 Conclusion

The SmartCHANGE project represents a significant step toward improving the early detection and prevention of non-communicable diseases (NCDs) in children and youth. While the tool presented in this paper is a demo version demonstrating some basic functionalities, our future work will focus on developing a more comprehensive web application for medical professionals and a mobile application for families. We also plan to enhance the tool

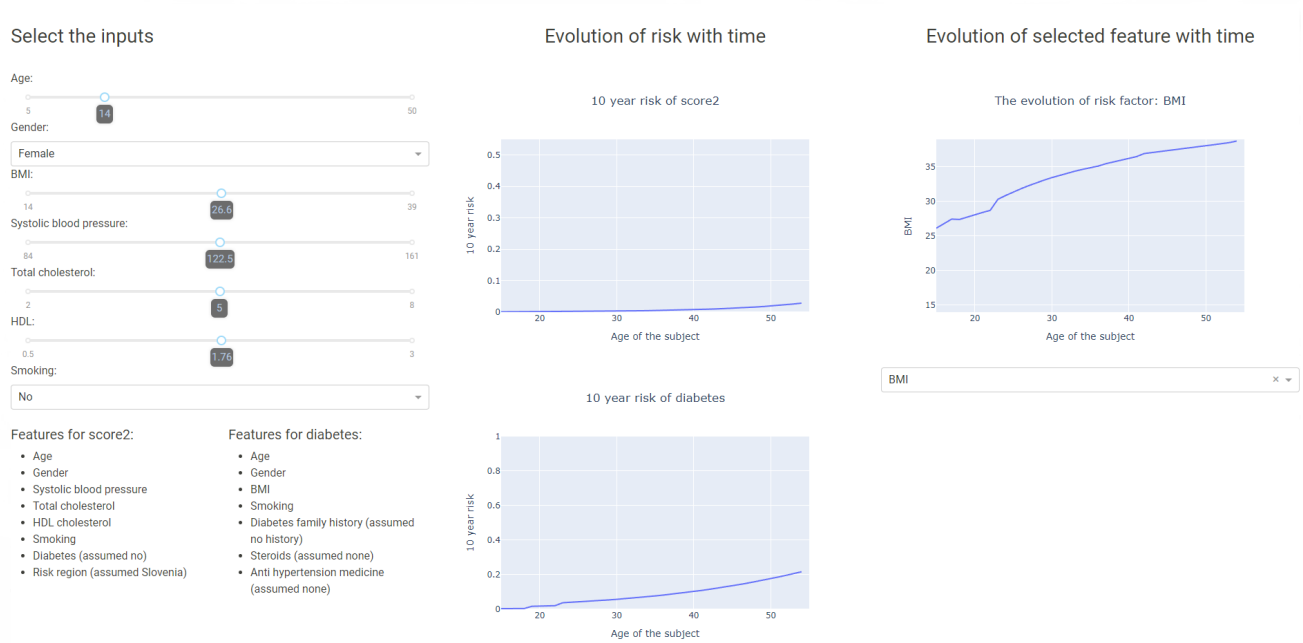


Figure 4: The figure is a dashboard interface that allows users to input various health-related parameters and observe the evolution of associated risks over time.

by replacing the current SCORE2 and QxMD risk models with more advanced models—Test2Prevent for diabetes and Healthy Heart Score for cardiovascular disease—incorporating features related to diet and physical activity. Additionally, the application will be updated to meet medical experts' needs based on their feedback.

Acknowledgements

This work was carried out as a part of the SmartCHANGE project, which received funding from the European Union's Horizon Europe research and innovation program under grant agreement No 101080965. The SLOfit dataset for wasvided by the University of Ljubljana (courtesy of Gregor Jurak et al.), the LGS dataset was provided by KU Leuven (courtesy of Martine ThomThomashe AFINA-TE dataset was provided by the University of Porto (courtesy of José Ribeiro) and the the University of Turku provided the YFS dataset are grateful for their support.

References

[1] P. S. Azzopardi, S. J. C. Hearps, K. L. Francis, E. C. Kennedy, A. H. Mokdad, N. J. Kassebaum, S. Lim, and et al. 2019. Progress in adolescent health and wellbeing: tracking 12 headline indicators for 195 countries and territories, 1990-2016. *Lancet*, 393, 10190, (Mar. 2019), 1101–1120.

[2] Gaston P Beunen, Robert M Malina, Marc A Van't Hof, Jan Simons, Michel Ostyn, Roland Renson, and Dirk Van Gerven. 1988. *Adolescent growth and motor performance: A longitudinal study of Belgian boys*. Human Kinetics Publishers.

[3] SmartCHANGE Consortium. 2024. Smartchange - horizon europe project. Accessed: 2024-09-02. (2024). https://www.smart-change.eu/.

[4] K. Corder, E. M. van Sluijs, I. Goodyer, C. L. Ridgway, R. M. Steele, D. Bamber, V. Dunn, S. J. Griffin, and U. Ekelund. 2011. Physical activity awareness of british adolescents. *Archives of Pediatrics Adolescent Medicine*, 165, 3, 281–289.

[5] A. García-Hermoso, R. Ramírez-Campillo, and M. Izquierdo. 2019. Is muscular fitness associated with future health benefits in children and adolescents? a systematic review and meta-analysis of longitudinal studies. *Sports Medicine*, 49, 7, (July 2019), 975–989.

[6] D. C. Jr Goff, D. M. Lloyd-Jones, G. Bennett, S. Coady, R. B. D'Agostino, and R. Gibbons. 2014. American college of cardiology/american heart association task force on practice guidelines. 2013 acc/aha guideline on the assessment of cardiovascular risk: a report of the american college of cardiology/american heart association task force on practice guidelines. *Circulation*, 129, 25, (June 2014), S49–S73.

[7] Noelia González-Gálvez, Jose Carlos Ribeiro, and Jorge Mota. 2022. Cardiorespiratory fitness, obesity and physical activity in schoolchildren: the effect of mediation. *International journal of environmental research and public health*, 19, 23, 16262–16270. ObjectType-Article-1. DOI: 10.3390/ijerph192316262.

[8] S. J. Griffin, P. S. Little, C. N. Hales, A. L. Kinmonth, and N. J. Wareham. 2000. Diabetes risk score: towards earlier detection of type 2 diabetes in general practice. *Diabetes/Metabolism Research and Reviews*, 16, 3, 164–171.

[9] D. R. Jacobs, J. G. Woo, A. R. Sinaiko, S. R. Daniels, J. Ikonen, and M. Juonala. 2022. Childhood cardiovascular risk factors and adult cardiovascular events. *New England Journal of Medicine*, 386, 19, (May 2022), 1765–1777.

[10] Markus Juonala et al. 2008. Cohort profile: the cardiovascular risk in young finns study. *International Journal of Epidemiology*, 37, 6, 1220–1226.

[11] Gregor Jurak et al. 2020. Slofit surveillance system of somatic and motor development of children and adolescents: upgrading the slovenian sports educational chart. *Acta Universitatis Carolinae. Kinanthropologica*, 56, 1, 28–40. DOI: 10.14712/23366052.2020.4.

[12] H. C. Jr McGill, C. A. McMahan, E. E. Herderick, G. T. Malcom, R. E. Tracy, and J. P. Strong. 2000. Origin of atherosclerosis in childhood and adolescence. *American Journal of Clinical Nutrition*, 72, 5, (Nov. 2000), 1307S–1315S.

[13] K. Pahkala, H. Hietalampi, T. T. Laitinen, J. S. Viikari, T. Rönnemaa, H. Niinikoski, and et al. 2013. Ideal cardiovascular health in adolescence: effect of lifestyle intervention and association with vascular intima-media thickness and elasticity (the special turku coronary risk factor intervention project for children [strip] study). *Circulation*, 127, 18, (May 2013), 2088–2096.

[14] J. R. Ruiz, I. Cavero-Redondo, F. B. Ortega, G. J. Welk, L. B. Andersen, and V. Martinez-Vizcaino. 2016. Cardiorespiratory fitness cut points to avoid cardiovascular disease risk in children and adolescents; what level of fitness should raise a red flag? a systematic review and meta-analysis. *British Journal of Sports Medicine*, 50, 13, 773–779.

[15] T. J. Saunders, C. E. Gray, V. J. Poitras, J. P. Chaput, I. Janssen, P. T. Katzmarzyk, and et al. 2016. Combinations of physical activity, sedentary behaviour and sleep: relationships with health indicators in school-aged children and youth. *Applied Physiology, Nutrition, and Metabolism*, 41, 6, (June 2016), 486–505.

[16] Johan Simons, Gaston Beunen, Roland Renson, Albrecht L. M. Claessens, Bernard Vanreusel, and Jos A. V. Lefevre. 1990. *Growth and fitness of Flemish girls: The Leuven Growth Study*. Human Kinetics, Champaign, IL.

[17] SCORE2 working group and ESC Cardiovascular risk collaboration. 2021. SCORE2 risk prediction algorithms: new models to estimate 10-year risk of cardiovascular disease in Europe. *European Heart Journal*, 42, 25, (June 2021), 2439–2454.

[18] World Health Organization. 2018. Global Health Estimate 2016: Deaths by Cause, Age, Sex, by Country and by Region, 2000-2016. World Health Organization.