SmartCHANGE Risk Prediction Tool: Next-Generation Risk Assessment for Children and Youth

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

Marko Jordan, Sebastjan Kramar, Ana Krstevska, Marcel Založnik
Jožef Stefan Institute,
Jožef Stefan International
Postgraduate School,
Ljubljana, Slovenia

Lotte van der Jagt
Harm op den Akker
Martijn Vastenburg
Research & Development
ConnectedCare
Nijmegen, The Netherlands

Valentina Di Giacomo
Elena Mancuso
Engineering Ingegneria Informatica SpA
Rome, Italy

Dario Fenoglio
Gabriele Dominici
Università della Svizzera italiana
Lugano, Switzerland

Mitja Luštrek
Jožef Stefan Institute,
Jožef Stefan International
Postgraduate School,
Ljubljana, Slovenia

Abstract

Non-communicable chronic diseases (NCDs), largely driven by lifestyle factors such as poor nutrition, physical inactivity, and obesity, account for over 70% of mortality in Europe. While prevention has traditionally focused on adults, growing evidence highlights the value of early intervention during childhood and adolescence to establish healthy behaviours and reduce long-term risk. This paper presents the updated SmartCHANGE platform, which harmonizes heterogeneous datasets, addresses missing information through synthetic data generation, and forecasts risk factors from childhood to adulthood. Forecasts are then applied to established cardiovascular and diabetes risk models, enabling long-term risk assessment. To ensure privacy, the platform incorporates federated learning for secure model training across distributed datasets. By combining synthetically generated data, predictive modelling, privacy-preserving infrastructure, and end-user applications, the updated SmartCHANGE platform supports early identification of at-risk youth and enables targeted, data-driven interventions to help reduce the future burden of NCDs.

Keywords

non-communicable diseases, risk prediction, synthetic data generation, federated learning, preventive healthcare

1 Introduction

Non-communicable diseases (NCDs), including cardiovascular disease and diabetes, cause over 70% of deaths in Europe [6]. Their onset is shaped by modifiable risk factors such as diet, physical inactivity, obesity, smoking, and alcohol use. While prevention strategies typically target adults, growing evidence highlights childhood and adolescence as critical periods for establishing lifelong health behaviours [5]. Addressing risk early can delay or prevent NCD onset and promote long-term well-being.

In this paper, we described an updated pipeline for predicting NCD risk in young people, building on our previous paper [4]. The new version introduces three advances: (i) broader harmonization of European cohort datasets through refined syntactic and semantic alignment; (ii) improved synthetic data generation that addresses heterogeneity of the datasets; and (iii) evaluation of advanced RNN-based architectures alongside conventional ML models. While the pipeline in the previous paper powered a simple demo, this one is integrated into the SmartCHANGE prototype that enables early identification of at-risk youth and supports the development of tailored preventive strategies. By combining harmonized datasets, predictive modelling, and privacy-preserving methods, it represents a step toward proactive, data-driven public health focused on youth as a critical stage for prevention. In addition, explainable AI was used to generate counterfactuals that support understanding of risk factors, and both web and mobile applications were developed to deliver these insights directly to healthcare professionals, adolescents, and families.

2 Baseline Predictive Approach

The models for forecasting risk factors are trained on seven heterogeneous datasets, none of which contain all the variables needed for risk prediction. The baseline predictive approach includes synthetic data generation and forecasting of individual risk factors from young to older age using various established machine-learning models. These forecast risk factors are then fed into established risk-prediction models to estimate the risk of cardiovascular disease and diabetes.

2.1 Synthetic Data Generation

The synthetic data generation was used to improve data completeness, enhance cross-dataset comparability, and support more robust forecasting and predictive modeling.

2.1.1 Generation of Diet Scores. The risk models required full dietary information, but none of the project datasets contained all the variables needed for diet scores. We therefore used the EUMenu dataset, which includes the complete set of dietary variables. Scores were first calculated for all EUMenu individuals. For project datasets with overlapping dietary or related features, we trained predictive models on EUMenu using only shared variables and generated synthetic diet scores accordingly. Given the task's simplicity and data structure, linear models were applied.

Information Society 2025, Ljubljana, Slovenia

https://doi.org/10.70314/is.2025.skui.7226

2.1.2 Generation of Other Data. We generated synthetic values for missing variables by constructing targeted sub-datasets and generating data with supervised learning. Each sub-dataset required core demographics (sex, age, weight, height); rows missing these were discarded to ensure stable baselines. A greedy search selected predictor sets that maximized coverage of missing entries, informativeness beyond demographics, and training sample size. Candidate sets were ranked by Score $= U \times V \times \sqrt{K}$, where U is the number of missing instances covered, V the number of predictors, and K the number of training rows.

For each sub-dataset, Gradient Boosting, Random Forest, and Linear Regression models were trained with k-fold cross-validation and grid search. Validation was assessed with Root Relative Squared Error (RRSE; where RRSE = 0 for perfect predictions, RRSE = 1 for baseline), and the best model generated the missing values. Overlaps were resolved by keeping predictions from the model with a lower RRSE. This process was repeated across variables to expand coverage while minimizing error. Data generation proceeded iteratively: after each pass, synthetic variables were evaluated with RRSE. Variables below a threshold were accepted and treated as ground truth in the next pass, with sub-datasets and models recomputed accordingly. The procedure terminated once no further variables met inclusion or performance plateaued, yielding a consistent. The mean RRSE of synthetic values in the final dataset was 0.795.

2.2 Risk Factor Forecasting

Having generated synthetic data, we constructed a merged dataset with no missing values. This dataset was used to train machine learning (ML) models to forecast health-related risk factors from childhood into adulthood. The predicted values were then applied as inputs to publicly available risk models to estimate the risk of developing NCDs.

We implemented a neural network (NN) with two dense layers (512 and 128 neurons) to capture non-linear patterns. Training used MSE loss, the Adam optimizer, ReLU activations, dropout (0.2), and early stopping. A single NN forecasted all risk factors simultaneously. Training and test data were prepared by generating all younger-to-older age pairs per individual. Inputs included gender, input and target age, and risk factors at the input age; targets were the same risk factors at the target age. This design enabled the model to learn age-progressive changes.

Input–output pairs were split into training, validation, and test sets, with each individual assigned to only one partition to avoid leakage. Stratification by dataset preserved source representation. Features were standardized with scikit-learn's StandardScaler. For comparison, we trained traditional ML models separately per variable: Linear Regression, Ridge Regression, Random Forest, and LightGBM (the latter via the lightgbm library). All models used default parameters and were trained/tested on the same pairs as the NN. Performance was measured with MAE and RRSE. Training used both real and synthetic data, but evaluation was restricted to real data. Input ages ranged from 6–18 years, and target ages from 18–55 years, matching the SmartCHANGE forecasting scope. The mean RRSE of the forecast values was 0.829.

2.3 Risk Models

We focused on two models: the Healthy Heart Score (HHS) for cardiovascular disease and Test2Prevent (T2P) for diabetes risk. Both include lifestyle factors such as physical activity and diet—essential for assessing younger populations and behavioural change—aligning with our goal of early prevention through modifiable risk factors. Using both models balanced clinical reliability with behavioural relevance, enabling a more comprehensive NCD risk assessment.

Our initial approach applied the models at age 55, the maximum forecastable age. This yielded inconsistent outputs: T2P produced 10-year risks (55–65), while HHS produced a 20-year risk (55–75). To resolve this, we instead reported cumulative risks to age 65, the most suitable endpoint given our data. Two strategies were evaluated: non-overlapping intervals and overlapping (hazard-averaging) intervals.

3 Advanced Unified Predictive Approaches

This section introduces advanced forecasting methods designed to work directly on heterogeneous datasets without requiring prior synthetic data generation. Despite their greater sophistication, their accuracy lags behind the more straightforward method that relies on synthetic data generation.

Synthetic data generation and forecasting are trained jointly within a single model, enabling the sharing of representations and feedback. Early layers provide initial estimates for both tasks, while later stages refine them by capturing complex temporal dependencies. Although SmartCHANGE uses only single-year inputs per user, the training dataset includes multi-year records, which reveal broader behavioural patterns.

Before entering the network, variables are normalized using training set statistics. Synthetic values are first generated in a linear block conditioned on age, gender, and BMI. This block consists of two fully connected layers (128 neurons + ReLU, then 21 neurons without activation). Forecasting then adds current age, future age, and gender, and predicts 21 risk factors across ages 6–55. The forecasting block differs by including an additional 128-neuron ReLU layer and more inputs. Forecasting is performed separately for each input year, and if multiple years exist, trajectories are averaged across target ages (e.g., data at 7, 9, and 12 yield three trajectories averaged per year).

This produces a time series of shape (50, 21). Appending masks for observed/synthetic values and gender gives (50, 43). Risk factor trajectories are then refined via a GRU block with bidirectional layers (128 or 21 hidden units) and a final 21-neuron linear layer. Predictions are finally de-normalized back to the original scale. The overall loss is the mean of two MAE terms: imputation and forecasting, with the latter computed only on ground-truth variables in the recorded output year.

The model was evaluated the same way as the one in Section 2.2, with the mean RRSE being 0.907. This is less than the RRSE from Section 2.2, indicating the need for further refinement of the unified approach.

4 Privacy Preservation and Explainability

Privacy Preservation. Within the SmartCHANGE project, health datasets are distributed across multiple countries and institutions. These sensitive data fall under strict regulations (e.g., GDPR), which prohibit cross-border sharing, and new pilot data remain stored locally, reinforcing isolation. Federated Learning (FL) addresses this by enabling collaborative training without moving raw data [3]. Two main challenges arise in deployment: pronounced heterogeneity across sites and residual privacy risks, since shared gradients can still leak information. To mitigate these, we developed distribution-aware, privacy-preserving FL strategies tailored to real-world healthcare [2]. Instead of a single global model, our approach builds compact, differentially private descriptors of each client's data distribution, clustering similar clients to train specialized models. This improves robustness to variability and temporal drift while ensuring fairer predictions, including for underrepresented groups. On the privacy side, model partitioning and communication-efficient aggregation reduce leakage without heavy cryptography by fragmenting gradients and distributing aggregation. Together, these strategies enable scalable, robust, and privacy-preserving FL pipelines for health risk prediction.

Explainability. Beyond predictive accuracy, effective NCD risk assessment must also provide transparent explanations and actionable guidance. For this, we adapt the Counterfactual Concept Bottleneck Model (CF-CBM) [1] to early-life health data. Instead of relying on predefined concepts—often unavailable or inconsistently annotated—our model learns patient feature distributions via a variational autoencoder (VAE), ensuring the latent space captures key generative factors of early-life trajectories. Counterfactuals are then generated following CF-CBM principles: given a patient profile and its predicted risk, the system proposes minimally altered, realistic configurations that would change the outcome. For example, if a child is predicted at high diabetes risk, the model may suggest plausible counterfactual profiles where lifestyle or physiological factors are adjusted to reduce risk. By embedding counterfactual reasoning directly into the pipeline, this approach goes beyond post-hoc interpretability. It both explains which factors drive predictions and identifies how risk can be reduced, offering clinicians and families actionable, personalized strategies for early prevention.

5 Architecture and User Applications

Architecture. The SmartCHANGE platform (Figure 1) is a modular, microservices-based system for AI-driven health interventions in children and adolescents. It integrates the developed predictive pipeline described in the previous sections with secure, scalable, and privacy-preserving technologies, with emphasis on GDPR compliance and explainable AI. Two main client interfaces are provided: the HappyPlant mobile app for families and youth, and a web application for healthcare professionals (HCPs).

Authentication and authorization are handled through the OpenID Connect (OIDC) protocol, with role-based access control and single sign-on. Additional safeguards include encrypted communication, pseudonymization, and immutable audit logging. Together, the SmartCHANGE platform, HappyPlant, and the HCP web interface form an integrated ecosystem for preventive healthcare, uniting advanced technical architecture with user-centered design to deliver effective, scalable, and personalized interventions.

Web Application. The web application for HCPs serves as a clinical dashboard, enabling them to access patient data, assess long-term risk for metabolic diseases (currently diabetes and CVD, although it can be scaled to integrate additional prediction models), and support behaviour change strategies. The interface is structured around a clinically aligned workflow — Consultation, Assessment, and Intervention — mirroring real-world practices.

Mobile Application. While intelligent risk predictions support HCPs in guiding clients, evidence and co-creation results show that simply communicating risks is insufficient for sustainable behaviour change in adolescents and families. The HappyPlant app was designed to address this gap. Rather than focusing on risks, it adopts a playful plant-growth analogy: users care for a virtual plant by completing daily and weekly personalized challenges linked to long-term health goals set by the HCP. The app nudges users towards the most suitable challenges but leaves the final choice to them, supporting autonomy and agency.

To foster long-term engagement, fully grown plants can be placed in the user's Goal Garden, which both showcases past achievements and acts as a reinforcement mechanism. In today's reward-driven context, the Goal Garden also enables saving towards real-life rewards set by parents, further motivating users. The app's design emerged from an extensive co-creation process and iterative validation with users, who responded positively to the analogy, challenge, and reward structure, as well as the aesthetics. Development was kept flexible, with adjustments made to align the app with other SmartCHANGE components.

6 Conclusion

This paper provides a concise description of the SmartCHANGE pipeline, which integrates harmonized datasets, synthetic data generation, federated learning, and explainable AI into a secure platform for early NCD risk prediction and prevention. Through the HappyPlant app and professional interface, these methods are translated into user-centered interventions that support sustainable behaviour change in youth. Detailed descriptions of the individual components will be published separately.

Acknowledgements

This work was carried out as 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 and ACDsi datasets were provided by the University of Ljubljana (courtesy of Gregor Jurak et al.), the LGS dataset was provided by KU Leuven (courtesy of Martine Thomis), AFINA-TE dataset was provided by the University of Porto (courtesy of José Ribeiro), ABCD was provided by VUMC, HELENA dataset was provided by Helena study group (courtesy of Francisco Ortega), and the University of Turku provided the YFS dataset. We are grateful for their support.

References

[1] Gabriele Dominici, Pietro Barbiero, Francesco Giannini, Martin Gjoreski, Giuseppe Marra, and Marc Langheinrich. 2025. Counterfactual concept bottleneck models. In *The Thirteenth International Conference on Learning Representations*. https://openreview.net/forum?id=w7pMjyjsKN.

[2] Dario Fenoglio, Gabriele Dominici, Pietro Barbiero, Alberto Tonda, Martin Gjoreski, and Marc Langheinrich. 2024. Federated behavioural planes: explaining the evolution of client behaviour in federated learning. In *Advances in Neural Information Processing Systems (NeurIPS 2024), Vol. 37*, 112777–112813.

[3] Dario Fenoglio, Daniel Josifovski, Alessandro Gobbetti, Mattias Formo, Hristijan Gjoreski, Martin Gjoreski, and Marc Langheinrich. 2023. Federated learning for privacy-aware cognitive workload estimation. In *Proceedings of the 22nd International Conference on Mobile and Ubiquitous Multimedia (MUM '23)*. ACM, New York, NY, USA, 25–36. DOI: 10.1145/3626705.3627783.

[4] Marko Jordan, Nina Reščič, Sebastjan Kramar, Marcel Založnik, and Mitja Luštrek. 2024. Smartchange risk prediction tool: demonstrating risk assessment for children and youth. In *Slovenska konferenca o umetni inteligenci. Zvezek A: zbornik 27. mednarodne multikonference Informacijska družba - IS 2024 : 10.–11. oktober, Ljubljana, Slovenija = Slovenian Conference on Artificial Intelligence. Vol. A : proceedings of the 27th International Multiconference Information Society - IS 2024*. Ljubljana, Slovenia, 71–74.

[5] 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.

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

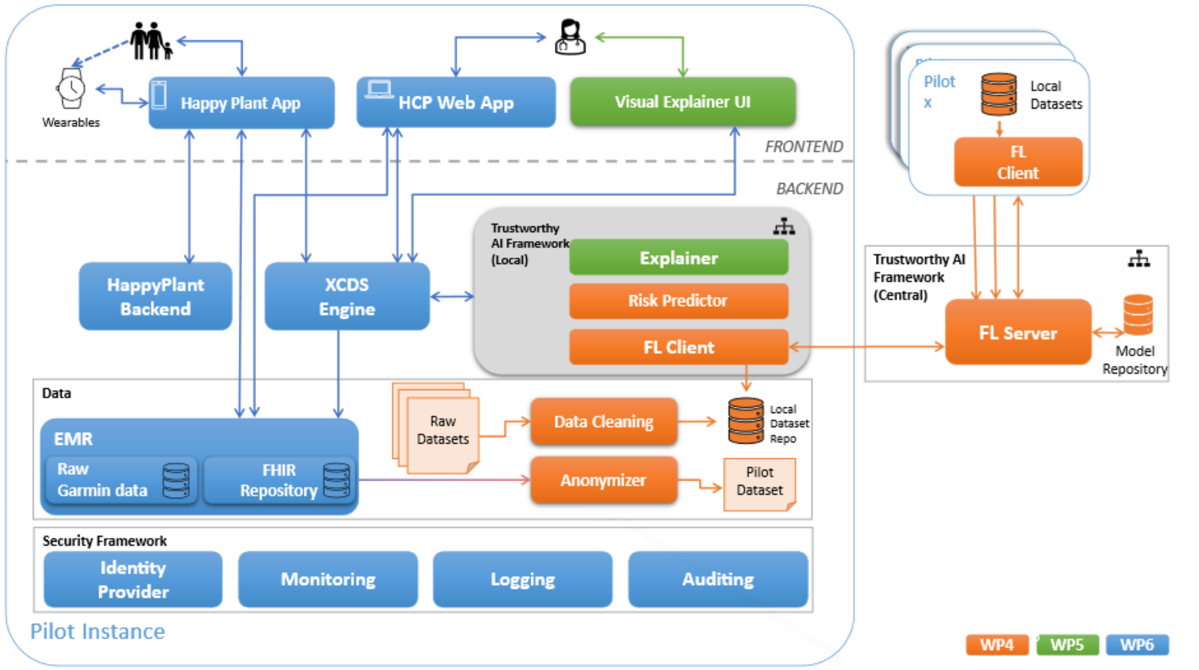

Figure 1: Logical Architecture of the SmartCHANGE Platform, including the mobile app (HappyPlant) and the web-app for healthcare professionals, connected to a central FHIR-compliant repository and featuring a Trustworthy AI Framework with federated learning, explainability, and secure communication via the XCDS Engine.

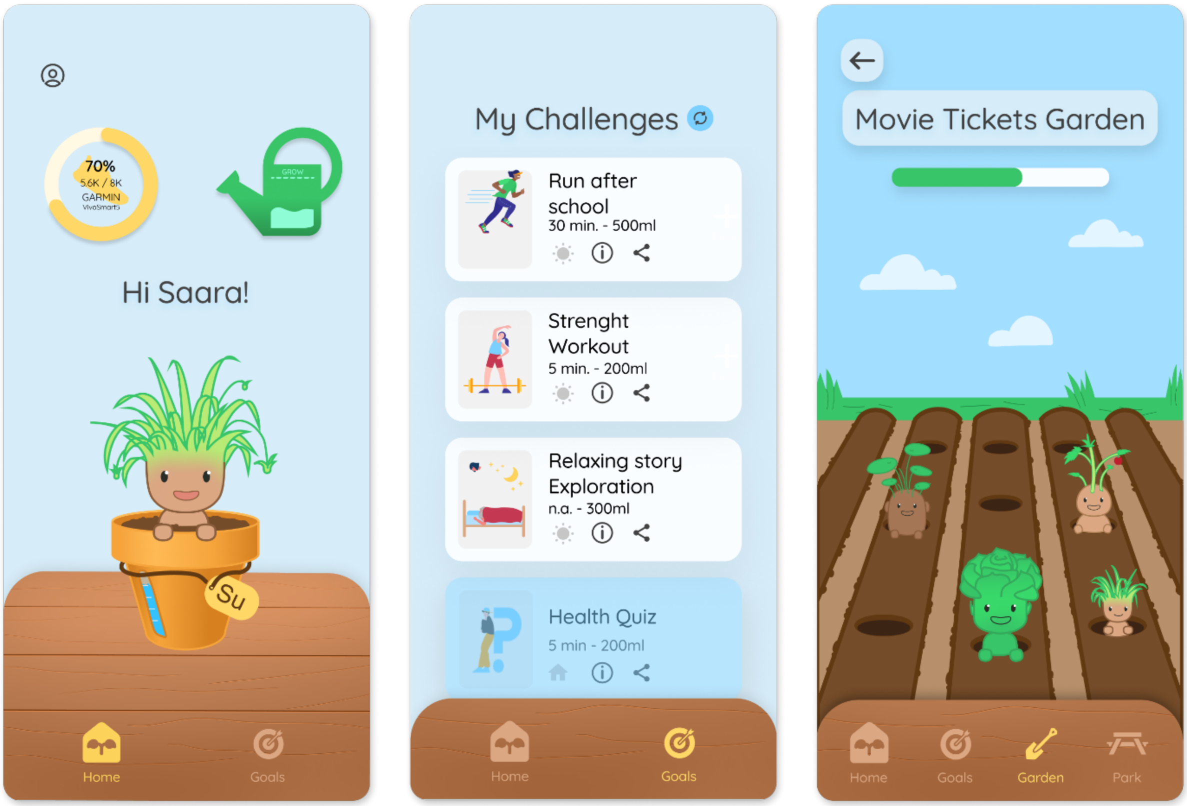

Figure 2: HappyPlant app screens: the home, challenge and goal garden screens.