

JYU DISSERTATIONS 754

---

**Ilkka Rautiainen**

# Prediction Methods for Assessing the Development of Individual Health Status

---



UNIVERSITY OF JYVÄSKYLÄ  
FACULTY OF INFORMATION  
TECHNOLOGY

JYU DISSERTATIONS 754

---

**Ilkka Rautiainen**

# **Prediction Methods for Assessing the Development of Individual Health Status**

Esitetään Jyväskylän yliopiston informaatioteknologian tiedekunnan suostumuksella  
julkisesti tarkastettavaksi Agoran auditoriossa 2  
maaliskuun 1. päivänä 2024 kello 12.

Academic dissertation to be publicly discussed, by permission of  
the Faculty of Information Technology of the University of Jyväskylä,  
in building Agora, auditorium 2, on March 1, 2024, at 12 o'clock.



JYVÄSKYLÄN YLIOPISTO  
UNIVERSITY OF JYVÄSKYLÄ

JYVÄSKYLÄ 2024

Editors

Marja-Leena Rantalainen

Faculty of Information Technology, University of Jyväskylä

Päivi Vuorio

Open Science Centre, University of Jyväskylä

Copyright © 2024, by the author and University of Jyväskylä

ISBN 978-951-39-9948-3 (PDF)

URN:ISBN:978-951-39-9948-3

ISSN 2489-9003

Permanent link to this publication: <http://urn.fi/URN:ISBN:978-951-39-9948-3>

## ABSTRACT

Rautiainen, Ilkka

Prediction methods for assessing the development of individual health status

Jyväskylä: University of Jyväskylä, 2024, 84 p. (+included articles)

There is growing interest in the application of artificial intelligence (AI) and machine learning in health-related tasks. Improved data utilization in health applications offers, in theory, multiple benefits. AI's evolution in health foresees it assisting and, in limited instances, replacing human judgment as well as enabling a shift from reactive treatments to proactive prevention. One of the potential targets concerns individual responses, making it possible to enhance personalized treatments. Improved preventive strategies targeting health issues not only reduce individual health risks but can also cut down the cost to society associated with medical care and productivity loss.

Health data's complexity, stemming from its incompleteness, variable quality, and diversity, poses obstacles in applying new methods. These challenges escalate during global data collection, hindered by diverse standards and practices.

This dissertation examines the potential of predictive modeling methods in different health-related applications, utilizing Finnish data from multiple sources. It includes three published articles and one manuscript, grouped into distinct use cases.

The first use case revolves around predicting obesity and overweight using childhood data. It consists of two articles, the Article I reviewing existing research on the field and the Article II evaluating the predictive capabilities of Finnish childhood growth data by applying various predictive modeling methods and contexts, benchmarked against prior studies.

The second use case in Article III focuses on predicting the development of 20-meter shuttle run test results (20MSRT) in adolescents. It is a unique study assessing the effectiveness of random forest in predicting 20MSRT development using an extensive dataset with diverse variables, including an exploration of the significance of individual variables, highlighting the need for a holistic view on health.

The third use case in Article IV describes a novel methodology for individual health assessment by creating a robust personal health index that addresses prevalent issues in structured health data, establishing a foundation to enhance various aspects of the rehabilitation process and beyond. The index enables a more holistic view of an individual's health status.

Keywords: predictive modeling, machine learning, health



## TIIVISTELMÄ (ABSTRACT IN FINNISH)

Rautiainen, Ilkka

Ennustemenetelmät yksilöllisen terveydentilan kehityksen arvioimiseen  
Jyväskylä: University of Jyväskylä, 2024, 84 s. (+artikkelit)

Tekoälyn ja koneoppimisen soveltaminen terveyteen liittyvissä tehtävissä on herättänyt paljon kiinnostusta. Edistyneempi aineistojen hyödyntäminen terveyssovelluksissa tarjoaa teoriassa monia etuja. Menetelmien kehittyminen mahdollistaa ihmisarvioinnin avustamisen tai jopa sen korvaamisen. Potentiaalia on myös siirtymisessä ongelmien reaktiivisesta hoidosta ennaltaehkäisyyn. Eräs mahdollisuus on keskittyä ns. yksilölliseen vasteeseen, mikä mahdollistaa mm. yksilöllisten hoitojen tehostamisen. Sen lisäksi, että tehostetut ennaltaehkäisevät strategiat vähentävät ihmisten terveysriskejä, ne voivat myös vähentää terveydenhuoltoon ja tuottavuuden menetyksiin liittyviä yhteiskunnan kustannuksia.

Terveysaineistojen monimutkaisuus, joka juontuu niiden puutteellisuudesta, vaihtelevasta laadusta ja monimuotoisuudesta, asettaa rajoitteita menetelmien soveltamiselle. Nämä haasteet kärjistyvät kun aineistoja kerätään ympäri maailman, sillä eri maissa on käytössä kirjavia standardeja ja käytäntöjä.

Väitöskirja tarkastelee koneoppimismenetelmien potentiaalia erilaisissa terveyteen liittyvissä sovelluksissa hyödyntämällä suomalaista dataa useista lähteistä. Se koostuu kolmesta julkaistusta artikkelista ja yhdestä käsikirjoituksesta, jotka on jaoteltu erillisiin käyttötapauksiin.

Ensimmäinen käyttötapaus keskittyy liikalihavuuden ja ylipainon ennustamiseen lapsuudenaikaisen aineiston avulla. Siihen kuuluu kaksi artikkelia. Artikkelit I käy läpi alan tutkimusta ja Artikkelit II arvioi suomalaisen lapsuuden kasvudatan ennustemahdollisuuksia soveltamalla erilaisia ennustemallinnuksen menetelmiä ja asetelmia sekä vertailee tuloksia aiempiin tutkimuksiin.

Toinen käyttötapaus Artikkelissa III keskittyy kardiorespiratorisen kunnon kehittymisen ennustamiseen nuorilla. Kyseessä on harvinaislaatuinen tutkimus, joka arvioi satunnaismetsäluokittelijan tehokkuutta ennustamisessa käyttäen laajaa aineistoa monipuolisilla muuttujilla. Myös yksittäisten muuttujien merkityksiä tutkitaan. Tulokset korostavat kokonaisvaltaisen terveysnäkökulman tarvetta.

Kolmannessa käyttötapauksessa Artikkelissa IV luodaan uusi menetelmä yksilöllisen terveyden arvioimiseksi luomalla henkilökohtainen terveysindeksi. Yleiset rakenteisten terveysaineistojen ongelmat on otettu huomioon. Kuvattu menetelmä luo pohjan eri osa-alueiden tehostamiseksi kuntoutusprosessissa ja sen ulkopuolella. Indeksillä mahdollistetaan kokonaisvaltaisemman yleiskuvan saamisen ihmisen terveydentilasta.

Avainsanat: ennustusmenetelmät, koneoppiminen, terveys

<b>Author</b>	Ilkka Rautiainen Faculty of Information Technology University of Jyväskylä Finland
<b>Supervisors</b>	Adjunct Professor Sami Äyrämö Faculty of Information Technology University of Jyväskylä Finland  Adjunct Professor Jukka-Pekka Kauppi Faculty of Information Technology University of Jyväskylä Finland  Adjunct Professor Toni Ruohonen Faculty of Information Technology University of Jyväskylä and Central Finland Wellbeing Service County Finland  Professor Pekka Neittaanmäki Faculty of Information Technology University of Jyväskylä Finland
<b>Reviewers</b>	Adjunct Professor Pekka Siirtola Faculty of Information Technology and Electrical Engineering University of Oulu Finland  Professor Mikko Tulppo Faculty of Medicine University of Oulu Finland
<b>Opponents</b>	Adjunct Professor Kari Kalliokoski Turku PET Centre University of Turku Finland  Professor Juha Röning Faculty of Information Technology and Electrical Engineering University of Oulu Finland

## ACKNOWLEDGEMENTS

This endeavor would not have been possible without my main supervisor, Sami. Your knowledge and wisdom gave me the confidence that I would be able to tackle any challenges associated with the dissertation process. Thank you for everything! I'm extremely grateful to Jukka-Pekka Kauppi. Your expertise and thoughtful comments truly made a difference. I'd also like to acknowledge the contribution of Toni Ruohonen. You and your networks were very valuable for this work. Special thanks to Pekka Neittaanmäki. Your projects provided the seeds that made this dissertation possible. Without your enthusiasm, this work might not even have started.

I am grateful to the two reviewers, Adjunct Professor Pekka Siirtola and Professor Mikko Tulppo. Their insightful comments significantly improved the dissertation. Special thanks to Adjunct Professor Kari Kalliokoski and Professor Juha Röning, who both kindly accepted the invitation to be my opponents.

I would like to express my deepest gratitude to the funding bodies of my research. The Jenny and Antti Wihuri Foundation provided funding for most of the research. The roots of the dissertation came from projects funded by Business Finland. Additionally, David Health Solutions played a significant role in providing funding. Additional funding was generously provided by the Faculty of Information Technology.

I want to extend my sincere thanks to all the co-authors and collaborators, whose domain knowledge and data made this research possible. Particularly, I'd like to mention Professor Urho Kujala who, in addition to his other valuable input, kindly agreed to be a referee for grant applications. Laura Joensuu provided essential contributions to this thesis, and through her, I was also able to be a part of the GenActive research group in the Faculty of Sport and Health Sciences. Many thanks to everyone in GenActive!

Special thanks to my previous office comrades, Susanne Jauhiainen and Lotta Palmberg, for inspiring discussions and peer support! I'd also like to acknowledge the people in the Digital Health Intelligence Laboratory and the Spectral Imaging Laboratory for making finalizing this thesis much more pleasant! Lastly, I'd like to recognize the staff in the Faculty of Information Technology and the Open Science Centre, the anonymous proofreader, and all the other people who have supported me during the work.

Many thanks to my multidisciplinary group of friends! During these years, we've had lunch meetings, traveled around Finland, hiked and biked, watched good (or very bad) movies, and engaged in other activities as well. These moments have been crucial for my well-being and have provided invaluable support. It's also heartening to witness fellow colleagues and friends navigating similar challenges during their own PhD studies over the last few years. Thank you Lauri Kortelainen, Elina Lähteelä, Juho Polet, Lauri Julkunen, Ida Vesterinen, Heidi Elmgren, Antti Moilanen, Jukka Ruokanen, Arttu Pekkarinen, Miio Seppänen, Minni Matikainen, Jaakko Vuori, Vieno Järventausta, Taneli Kupari-

nen, Henna Väyrynen, and Matti Lahdenmäki.

Kiitos äiti ja isä alusta lähtien antamastanne tuesta. Kiitos Olli hyvästä huonosta huumoristasi ja kaikesta muustakin. Katariina, on ollut upeaa tehdä tätä matkaa sinun kanssasi. Yhdessä tämä kaikki on ollut paljon helpompaa ja hauskeempaa. Kiitos rakas kaikesta!

Jyväskylä 12.2.2024

Ilkka Rautiainen

## LIST OF ACRONYMS

<b>20MSRT</b>	20-meter shuttle run test
<b>AGI</b>	Artificial general intelligence
<b>AI</b>	Artificial intelligence
<b>AUC</b>	Area under an ROC curve
<b>BO</b>	Bayesian (hyperparameter) optimization
<b>CPI</b>	Conditional permutation importance
<b>CRF</b>	Cardiorespiratory fitness
<b>CV</b>	Cross-validation
<b>EI</b>	Expected improvement
<b>FN</b>	False negatives
<b>FP</b>	False positives
<b>GP</b>	Gaussian process
<b>ICF</b>	International Classification of Functioning, Disability and Health
<i>k</i> NN	<i>k</i> -nearest neighbors
<b>LLM</b>	Large language model
<b>MCC</b>	Matthews correlation coefficient
<b>ML</b>	Machine learning
<b>OOB</b>	Out-of-bag
<b>PCA</b>	Principal component analysis
<b>PI</b>	Permutation importance
<b>RF</b>	Random forest
<b>ROC</b>	Receiver operating characteristics
<b>SVC</b>	Support vector classifier
<b>SVM</b>	Support vector machine
<b>TN</b>	True negatives
<b>TP</b>	True positives
<b>WHO</b>	World Health Organization

## LIST OF FIGURES

FIGURE 1	A confusion matrix for binary classification.....	22
FIGURE 2	An ROC curve example for binary classification .....	24
FIGURE 3	Model complexity and its generalization ability .....	30
FIGURE 4	Data split in a 5-fold cross-validation.....	31
FIGURE 5	A Gaussian process in Bayesian optimization .....	34
FIGURE 6	A linear SVC .....	39
FIGURE 7	An SVC with polynomial kernel.....	40
FIGURE 8	A decision tree.....	41
FIGURE 9	A random forest .....	43
FIGURE 10	A $k$ NN classifier .....	46

# CONTENTS

ABSTRACT

TIIVISTELMÄ (ABSTRACT IN FINNISH)

ACKNOWLEDGEMENTS

LIST OF ACRONYMS

LIST OF FIGURES

CONTENTS

LIST OF INCLUDED ARTICLES

1	INTRODUCTION .....	15
1.1	Research background.....	15
1.2	Principles of predictive modeling.....	17
1.3	Research questions .....	18
1.4	Thesis overview .....	19
2	KEY TERMINOLOGY AND TECHNIQUES .....	20
2.1	Mathematical notations.....	20
2.2	Variable types .....	21
2.3	Performance metrics .....	21
2.4	Preparing the data .....	25
2.5	Model selection and assessment.....	29
2.5.1	Cross-validation and bootstrap .....	30
2.6	Hyperparameter optimization .....	32
2.7	Explaining the predictions.....	35
3	PREDICTION METHODS.....	37
3.1	Linear and logistic regression .....	37
3.2	Support vector machine .....	38
3.3	Decision trees.....	40
3.4	Random forest.....	42
3.5	Other methods .....	45
4	HEALTH DOMAIN APPLICATIONS .....	47
4.1	Obesity and overweight .....	49
4.2	Cardiorespiratory fitness development .....	50
4.3	Health monitoring .....	51
4.4	Ethical considerations .....	54
5	SUMMARY OF THE INCLUDED ARTICLES.....	55
5.1	Article I: Predicting overweight and obesity in later life from childhood data: a review of predictive modeling approaches.....	55
5.2	Article II: Predicting future overweight and obesity from childhood growth data: a case study.....	56

5.3	Article III: Precision exercise medicine: predicting unfavourable status and development in the 20-m shuttle run test performance in adolescence with machine learning.....	57
5.4	Article IV: Utilizing the International Classification of Functioning, Disability and Health (ICF) in forming a personal health index	58
6	CONCLUSIONS .....	60
6.1	Potential in obesity and overweight prediction (Q1).....	60
6.2	Potential in cardiorespiratory fitness development prediction (Q2)	61
6.3	Developing an ICF-based personal health index and its influence on health assessments and ML (Q3).....	62
6.4	Limitations and future work.....	63
	YHTEENVETO (SUMMARY IN FINNISH) .....	65
	BIBLIOGRAPHY.....	68
	INCLUDED ARTICLES	



## LIST OF INCLUDED ARTICLES

- I Ilkka Rautiainen and Sami Äyrämö. Predicting overweight and obesity in later life from childhood data: A review of predictive modeling approaches. *Computational Sciences and Artificial Intelligence in Industry: New Digital Technologies for Solving Future Societal and Economical Challenges*, pages 203–220, 2021.
- II Ilkka Rautiainen, Jukka-Pekka Kauppi, Toni Ruohonen, Eero Karhu, Keijo Lukkarinen, and Sami Äyrämö. Predicting future overweight and obesity from childhood growth data: A case study. *Computational Sciences and Artificial Intelligence in Industry: New Digital Technologies for Solving Future Societal and Economical Challenges*, pages 189–201, 2021.
- III Ilkka Rautiainen, Laura Joensuu, Sami Äyrämö, Heidi J Syväoja, Jukka-Pekka Kauppi, Urho M Kujala, and Tuija H Tammelin. Precision exercise medicine: predicting unfavourable status and development in the 20-m shuttle run test performance in adolescence with machine learning. *BMJ Open Sport & Exercise Medicine*, 7:e001053, 2021.
- IV Ilkka Rautiainen, Lauri Parviainen, Veera Jakoaho, Sami Äyrämö, and Jukka-Pekka Kauppi. Utilizing the International Classification of Functioning, Disability and Health (ICF) in forming a personal health index. *Manuscript*, 2023.

The author of this thesis is the first author in all of the included articles. The first authorship of Article III is shared with Laura Joensuu. In all articles, the author was mainly responsible for all tasks, excluding data collection and curation. These tasks included writing the original drafts of the articles, designing the methodology, implementing the methods, and validating the results.



# 1 INTRODUCTION

This thesis delves into the application of predictive modeling methods to health data, along with the introduction of a novel approach for evaluating health. Each of the three case studies places an emphasis on individual health, providing insights into personalized healthcare strategies.

## 1.1 Research background

The utilization of existing health data can be significantly optimized. Mooney and Pejaver (2018) highlighted that the diverse array of big data in public health can unveil insights into questions that were once beyond our reach. This influx of data opens the door to exploring and validating questions that may not have previously been contemplated (Dhar 2013). Moreover, such novel methodologies pave the way for the generation of new hypotheses (Ludwig and Mullainathan 2023).

In the foreseeable future, **artificial intelligence** (AI) has the potential to assist physicians in making better clinical decisions and, in some limited cases, even replace human judgment (Jiang et al. 2017). These and other advancements in technology have the potential to move health care more from the reactive treatment of already observed health issues to a more proactive approach that aims to prevent problems even before they emerge (Schiavone and Ferretti 2021; Waldman and Terzic 2019). One potential area would be to enhance personalized treatment by using methods in **machine learning** (ML) and **predictive modeling** (Alyass, Turcotte, and Meyre 2015; Iniesta, Stahl, and McGuffin 2016). In essence, ML is recognized as a specialized branch within the broader field of AI (Alpaydm 2014). On the other hand, the concept of predictive modeling, which is discussed later, is characterized by a more precise and narrowly tailored definition.

A significant share of social welfare and health-care expenditure is incurred by two demographics—namely small children and the elderly (Hujanen et al. 2008). Elevating the success rate of interventions among these groups could not

only expedite recovery and preempt future complications but also yield societal benefits by curtailing overall costs.

For instance, Reini and Honkatukia (2016) discovered that enhancing early detection of type 2 diabetes could contribute a two billion euro boost to Finland's gross domestic product by reducing sickness-related absences. Another example of a field that might benefit from new preventive strategies is overweight and obesity prevention. Obesity's correlation with a spectrum of detrimental health outcomes, including cardiovascular diseases, type 2 diabetes, and cancer, is well-documented (Shrestha et al. 2016). Additionally, the economic implications are considerable, with obese individuals incurring higher health-care costs and exhibiting more absences from work and reduced productivity (Kleinman et al. 2014).

This thesis primarily addresses the challenge of prediction, with the insights derived being applicable in both prevention and rehabilitation, and potentially further. Foremost, it is crucial to establish methodologies for the early identification of individuals at risk of impending health complications (Ruohonen et al. 2018). Such proactive detection not only facilitates timely interventions but also yields considerable economic advantages, as evidenced by the work of Reini and Honkatukia (2016).

Conversely, when a health condition is already present, the rehabilitation process must be expedited and optimized for efficiency. Rehabilitation encompasses a variety of physical modalities and therapeutic exercises (Espregueira-Mendes, Barbosa Pereira, and Monteiro 2011). Yet, as Filos et al. (2017) pointed out, there has been a lack of focus on tailoring beneficial performance parameters for exercises at a personalized level. They also concluded that predictive modeling can significantly enhance personalized exercise guidance, thereby streamlining rehabilitation.

Nonetheless, the direct application of the vast array of health-related data is fraught with challenges, including issues of incompleteness, data quality, and heterogeneity (Dinov 2016; Ghassemi et al. 2020; Viceconti, Hunter, and Hose 2015). Data heterogeneity, in particular, refers to data diversity, such as in data types, file formats, encoding methods, and semantic discrepancies (L'Heureux et al. 2017). These factors collectively impede the efficient harnessing of large datasets for AI advancements (Krumholz 2014; L'Heureux et al. 2017).

These challenges are magnified further when considering the integration of data collected on a global scale from disparate sources. The diversity in standards, practices, cultural norms, survey methodologies, and languages across nations renders the global standardization of measurement and treatment protocols a near-impossible endeavor.

Given the widespread availability of ML algorithms in programming libraries such as scikit-learn (Pedregosa et al. 2011), TensorFlow (Abadi et al. 2015), Keras (Chollet et al. 2015), and Pytorch (Paszke et al. 2019), the primary focus of and challenge in predictive modeling applications lie in the realm of data. The preparation of data for each unique case remains a highly task-specific endeavor, the full automation of which is inherently challenging (Brownlee 2020). As a

result, progress in health data preparation techniques significantly contributes to enhancing the efficacy of predictive modeling applications, offering valuable steps toward more informed and accurate analyses in the field.

## 1.2 Principles of predictive modeling

The interrelated fields of ML, data mining, predictive analytics, and data science are frequently mentioned in tandem. These disciplines often intersect, blurring the lines that distinguish one from another. In this thesis, the focus is on predictive modeling, an approach that encompasses the entire process from defining objectives to designing studies and collecting data.

Breiman (2001b) and Shmueli (2010) discussed the distinctions between **explanatory** and **predictive** types of statistical modeling. The term "modeling" here refers to the entire process, from setting goals to designing research protocols and collecting data. Predictive modeling is further categorized based on the nature of the response variable. **Classification** refers to predicting categorical responses, while **regression** deals with continuous responses (Tan et al. 2013).

Consider  $x$  as a vector of **input variables**, also known as **predictors**. Nature can be seen as associating these variables with **outputs**, or **response variables**, represented by  $y$ . This association can be viewed as a "black box", containing unknown processes (Breiman 2001b; Callahan and Shah 2017). The goal is to use  $x$  to predict  $y$  for each observation, a process called supervised learning (Hastie, Tibshirani, and Friedman 2009). In contrast, unsupervised learning lacks known response variables for guidance (Hastie, Tibshirani, and Friedman 2009). **Reinforcement learning** is another category, where outputs are actions aimed at achieving a goal efficiently, which is useful in applications such as robot navigation and gaming (Alpaydm 2014).

Breiman (2001b) described how explanatory modeling aims to fill the black box of the data generation process by constructing a model that best fits the given data. Shmueli (2010) characterized explanatory modeling as the application of statistical models to data for the purpose of testing causal hypotheses. In contrast, predictive modeling has been described as the application of a statistical model or data mining algorithm to data with the aim of predicting new or future observations (Shmueli 2010). This delineation between the two modeling approaches is a crucial distinction in the field.

Breiman (2001b) and Shmueli (2010) contended that goodness-of-fit tests and related techniques fall short in assessing the suitability of a model for predictive tasks. Breiman (2001b) further asserted that predictive modeling methods are more capable in producing information about the structure of the relationship between inputs and responses.

In this work, the predictive modeling approach is favored over explanatory modeling because our primary goal is prediction, assessing how well current data can be extrapolated to future events. Prediction inherently does not necessitate an

understanding of the underlying processes; rather, it centers on pinpointing the most effective strategies, such as determining the ideal treatment option (Bzdok, Altman, and Krzywinski 2018). Consequently, there is no requirement for pre-existing causal hypotheses regarding data variables' interrelations. The data and model serve as tools for uncovering potentially novel insights that warrant further investigation. This approach of generating hypotheses through data-driven predictive modeling holds the promise of uncovering new insights, since it is predicated on fewer a priori assumptions about the subject matter (Oquendo et al. 2012). Nonetheless, the expertise of domain specialists is essential for interpreting findings and identifying those meriting additional study. Clearly, predictive models developed in research must undergo thorough evaluation and scrutiny by domain experts before they can be fully leveraged in practical applications (Khoury and Ioannidis 2014).

An additional advantage of predictive modeling is that it allows for a data-driven approach to variable selection, meaning variables can be chosen for a model using automated procedures. This also enables the consideration of a significantly larger set of initial candidate variables. In contrast, explanatory modeling is inherently hypothesis-driven, and automated variable selection is generally discouraged (Sainani 2014).

The predictive, data-driven approach to model building is not entirely free from subjective decisions (Mooney and Pejaver 2018). Common pitfalls in predictive modeling include issues with data preparation, validation problems such as data leakage, applying the model to data not encountered during training, and overfitting the model to the training data (Jauhiainen 2023; Kuhn and Johnson 2016).

Traditionally, health-care treatments have been prescribed based on their average efficacy across certain populations, under the assumption that what works for the majority will work for all (MacEachern and Forkert 2021; Tuena et al. 2020). However, individual responses to treatment can vary due to factors such as genetics (Roden 2016). **Precision medicine**, also known as **precision health**, challenges this one-size-fits-all approach by considering each person's unique characteristics (MacEachern and Forkert 2021; Tuena et al. 2020). An extension of this concept is **predictive precision medicine**, which involves predicting an individual's disease trajectory and response to treatment using a combination of biomarkers and personal data, including lifestyle and environmental factors (Tuena et al. 2020). Similarly, **precision exercise medicine** recognizes individual differences in how people respond to various exercise doses (Ross et al. 2019).

### 1.3 Research questions

The following are main research questions of this thesis:

- Q1: What is the potential of predictive modeling methods in the prediction of obesity and overweight in children and adolescents?

- Q2: What is the potential of predictive modeling methods in predicting cardiorespiratory fitness development in adolescents?
- Q3: Considering the challenges in structured health data, how can the adoption of the WHO's International Classification of Functioning, Disability, and Health (ICF) framework be utilized in the development of a comprehensive personal health index, and what impact could this index have on health assessments and the application of ML algorithms?

## 1.4 Thesis overview

This thesis centers on three distinct use cases. The initial use case delves into the prediction of overweight and obesity, featuring a literature review (Article I) and a case study focused on the body mass index (BMI) trajectories of Finnish children (Article II). The second use case assesses the application of predictive modeling to the progression of individual performance in the 20-meter shuttle run test (20MSRT), a common field test for cardiorespiratory fitness (CRF), during adolescence (Article III). The final use case is dedicated to devising a comprehensive personal health index within the context of rehabilitation (Article IV).

The structure of the thesis is as follows. Chapter 2 introduces key terminology and techniques essential to predictive modeling. Chapter 3 elucidates the specific prediction methods employed. Chapter 4 provides an overview of recent implementations of predictive modeling in the health domain. Chapter 5 offers a summary of the articles incorporated into the thesis. Finally, Chapter 6 contemplates the findings and discusses prospective implications of the research.

## 2 KEY TERMINOLOGY AND TECHNIQUES

This chapter discusses the key terminology and an array of techniques commonly utilized in predictive modeling. Topics addressed include essential mathematical notations (Section 2.1), variable types (Section 2.2), performance metrics (Section 2.3), data preparation procedures (Section 2.4), model selection and assessment (Section 2.5), hyperparameter optimization (Section 2.6), and methods for interpreting model predictions (Section 2.7).

### 2.1 Mathematical notations

A data matrix  $\mathbf{X}$ , representing  $N$  observations and  $p$  variables, is formulated as:

$$\mathbf{X} = \begin{bmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{bmatrix} = \begin{bmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N,1} & x_{N,2} & \cdots & x_{N,p} \end{bmatrix}. \quad (1)$$

Here,  $\mathbf{x}_i$  signifies a column vector and  $\mathbf{x}_i^T$  its transpose (a row of matrix  $\mathbf{X}$ ), while each  $x_{i,j}$  represents the scalar value for the  $i^{\text{th}}$  observation's  $j^{\text{th}}$  variable. For example,  $\boldsymbol{\beta}^T \mathbf{x}_i$  represents the dot product of vectors  $\boldsymbol{\beta}$  and  $\mathbf{x}_i$ . Additionally,  $\mathbf{x}_{\cdot,j}$  denotes a vector with all values of the  $j^{\text{th}}$  variable in a data sample  $\mathbf{X}$ . Furthermore, vectors are represented by bolded lowercase letters (e.g.,  $\mathbf{x}$ ,  $\boldsymbol{\lambda}$ , and  $\boldsymbol{\beta}$ ), while scalars are denoted by regular lowercase letters (e.g.,  $x$  and  $y$ ).

Expanding on  $\mathbf{X}$ , a dataset  $\mathbf{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$  includes the vectors illustrated in Eq. 1, along with a corresponding outcome  $y$  for each observation. Furthermore, the function  $\exp(\cdot)$  is equivalent to  $e^{(\cdot)}$ , and the set of real numbers is denoted by  $\mathbb{R}$ .



## 2.2 Variable types

**Structured data** is characterized by well-defined fields for each variable, which include numeric and/or textual information (Kantardzic 2020). Although **unstructured data**, such as video recordings and free-form medical reports, do exist in the health data domain, they are not discussed here, since they do not feature in the datasets used.

Variables in structured data fall into two primary categories—namely **categorical** and **numeric**. Categorical variables, also known as **qualitative** variables, are subdivided into **nominal** and **ordinal**, as follows (Kantardzic 2020; Tan et al. 2013):

- Nominal variables are typically represented by words and lack a meaningful order. Examples include diagnosed diseases or binary variables with two possible outcomes (Kantardzic 2020; Tan et al. 2013).
- Ordinal variables, or **ordered categorical** variables, are akin to nominal variables but possess a definable order. A common instance is a Likert scale response ranging from 1 to 5 (Hastie, Tibshirani, and Friedman 2009; Kantardzic 2020; Tan et al. 2013).

Numeric variables, also referred to as **quantitative** variables, are classified as either **interval scale** or **ratio scale**, as follows:

- Interval scale variables have an arbitrary zero point, such as temperatures in Celsius or calendar dates (Kantardzic 2020; Tan et al. 2013).
- Ratio scale variables feature an absolute zero point, with most numeric variables falling into this category, such as temperatures in Kelvin (Kantardzic 2020; Tan et al. 2013).

Additionally, variables are often described as **discrete** or **continuous**. Discrete variables have a finite or countably infinite set of values and can be associated with any of the four variable types mentioned (Kantardzic 2020; Tan et al. 2013). An example is the number of laps in the 20MSRT, which is a discrete variable on a ratio scale (Tomkinson et al. 2017). In contrast, continuous variables are represented by real numbers (Kantardzic 2020; Tan et al. 2013).

## 2.3 Performance metrics

Evaluating the performance of various models is crucial for comparing and determining their practical applicability. Positive cases are defined as instances belonging to the class of interest. For instance, in breast cancer diagnosis, a positive case would be labeled “malignant,” while a negative case would be “benign.”

**True positives** (TP) represent the count of positive cases accurately identified, **false positives** (FP) denote positive cases incorrectly identified, **true negatives** (TN) are negative cases accurately identified, and **false negatives** (FN) are negative cases incorrectly identified (Callahan and Shah 2017; Fawcett 2006). These outcomes are typically organized into a confusion matrix (see Figure 1).

True label	Malignant	57	9
	Benign	6	116
		Malignant	Benign
		Predicted label	

FIGURE 1 The confusion matrix for binary classification, applied to breast cancer data<sup>1</sup> using a  $k$ -nearest neighbors classifier (Pedregosa et al. 2011), illustrates the outcomes of the classification process. Within this matrix, the 57 malignant cases that were accurately identified as malignant represent the true positives, while the nine malignant cases that were mistakenly identified as benign are the false positives. Conversely, the 116 benign cases that were correctly classified as benign constitute the true negatives, and the six benign cases that were erroneously classified as malignant are the false negatives.

From the outcomes of TP, FP, TN, and FN, several performance metrics can be derived, with **accuracy**, **sensitivity**, and **specificity** being among the most commonly utilized. They are defined as

$$\text{accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}} \quad (2)$$

$$\text{sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}} \quad (3)$$

$$\text{specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad (4)$$

Accuracy (Eq. 2) represents the proportion of correctly classified observations. However, this metric may present an overly optimistic view in cases of class imbalance. For example, if only one percent of observations belong to the positive class, classifying all observations as negative would result in 99% accuracy (Kantardzic 2020). To obtain meaningful results for practical applications, additional metrics are necessary to evaluate a model's performance.

<sup>1</sup> The breast cancer dataset (Wolberg et al. 1995) serves as a recurring example throughout this thesis. Consisting of 569 samples, the dataset features 30 continuous variables derived from digitized images of breast masses. Each sample is annotated with class information, categorized as either malignant or benign.

Sensitivity (Eq. 3) and specificity (Eq. 4) are particularly crucial in the medical field. Sensitivity measures the likelihood of correctly predicting a positive outcome when the true outcome is positive (Hastie, Tibshirani, and Friedman 2009), while specificity measures the likelihood of correctly predicting a negative outcome when the true outcome is negative (Hastie, Tibshirani, and Friedman 2009). For instance, a model with 80% sensitivity and 60% specificity indicates that it correctly predicts 80% of disease cases, but only 60% of non-disease cases are accurately classified. Selecting the final model involves balancing these two metrics, among others.

Three additional metrics often used are **precision**, **negative predictive value (NPV)** and **F-measure**, defined as

$$\text{precision} = \frac{TP}{TP + FP} \quad (5)$$

$$\text{NPV} = \frac{TN}{TN + FN} \quad (6)$$

$$\text{F-measure} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}}. \quad (7)$$

Precision (Eq. 5) assesses the ratio of true positive observations to all positive predictions, while NPV (Eq. 6) measures the ratio of true negative observations to all negative predictions. The F-measure (Eq. 7) computes the harmonic mean of precision and sensitivity, effectively combining these two metrics (Zaki, Meira Jr, and Meira 2014).

The balance between sensitivity and specificity is depicted in a **receiver operating characteristics (ROC) curve**, which plots from point (0,0) to point (1,1), as illustrated in Figure 2. Constructing an ROC curve requires a classifier that provides continuous outputs, such as class prediction probabilities. The curve is formed by sequentially calculating sensitivity and specificity for various threshold values. The ROC curve's efficacy is summarized by the **area under the ROC curve (AUC)**,  $\mathbb{R} \in [0, 1]$ , with a higher AUC signifying better performance (Fawcett 2006). The AUC effectively merges sensitivity and specificity into a single metric and, unlike accuracy, provides a more reliable measure of classifier performance across both balanced and imbalanced datasets (Huang and Ling 2005).

While the AUC is a prevalent metric in research, critics (Chicco and Jurman 2023) have pointed out its tendency to yield overoptimistic results by not accounting for precision and negative predictive value. They have suggested the **Matthews correlation coefficient (MCC)** as an alternative for binary classification. MCC only awards a high score within its range of  $\mathbb{R} \in [-1, 1]$  if all four confusion matrix metrics—sensitivity, specificity, precision, and NPV—are substantial (Chicco and Jurman 2023). It is defined as

$$\text{MCC} = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}. \quad (8)$$

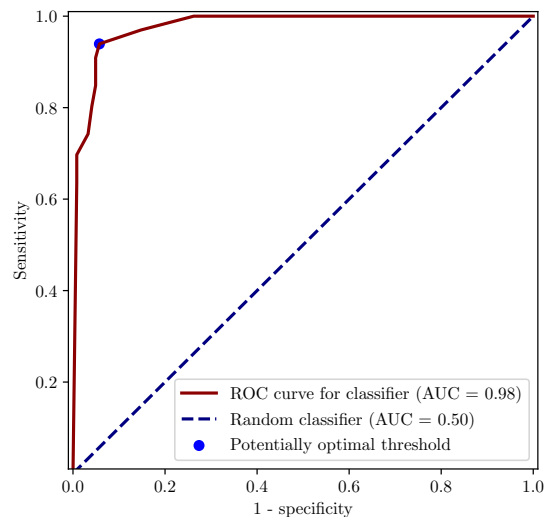


FIGURE 2 A receiver operating characteristics curve for binary classification using breast cancer data (Wolberg et al. 1995). At the point  $(0, 0)$ , all observations are classified as negative, resulting in a model sensitivity of 0%. Conversely, the specificity is 100%, since no positive cases have been falsely identified as negative. As the threshold is changed, the classifier begins to identify some positive cases. Theoretically, the optimal operating point is  $(0, 1)$ , indicating no misclassifications, but this is unachievable in practice. At this point, the sensitivity and specificity reach their theoretical maximum values. The optimal operating point is therefore as close as possible to point  $(0, 1)$ . At one such optimal point marked on the curve, the classifier achieves 93.9% sensitivity and 94.2% specificity, with an area under the curve (AUC) of 0.98. Random guesses would yield a straight line from  $(0, 0)$  to  $(1, 1)$ , equating to an AUC of 0.5. The example is adapted from documentation by Pedregosa et al. (2011).

A data-driven approach can determine a risk threshold that minimizes incorrect classifications, but this may not always be the most suitable threshold in practice (Wynants et al. 2019). The selection of the final threshold should weigh the benefits of correct classifications against the costs of incorrect ones, especially in clinical settings (Wynants et al. 2019). One strategy involves making use of domain-specific knowledge and individually determining the cost for each classification type—TP, FP, TN, and FN—then minimizing the expected total cost (Wynants et al. 2019).

The metrics above can also be extended for multi-class problems. One approach is to evaluate one class against all others. For example, in case of a three-class problem, there would be three class-specific sensitivity metric values. The first value is for sensitivity when class 1 is evaluated against classes 2 and 3, the second is for sensitivity when class 2 is evaluated against classes 1 and 3, and the third is for sensitivity when class 3 is evaluated against classes 1 and 2 (Zaki, Meira Jr, and Meira 2014). An overall metric for the classifier can be calculated by averaging these class-specific metrics (Zaki, Meira Jr, and Meira 2014). Alternatively, balanced metrics can be computed, with weightings derived from the

frequency of each class in the dataset (Grandini, Bagli, and Visani 2020).

Statistical tests are valuable for assessing whether models produce systematically good metrics. The metrics derived from models may exhibit significant random variation due to the division of data into training, testing, and validation sets as well as variations in the initial parameters of the models (Äyrämö, Kärkkäinen, and Majava 2007). For example, a one-sample  $t$ -test can be employed to determine whether the mean of multiple values, such as AUC values, significantly differs from a specified value, denoted as  $\mu_0$ . For instance, when testing with multiple AUC values to determine whether the model is better than random,  $\mu_0$  would be set as 0.5. The null hypothesis,  $H_0$ , posits that the mean from the sample is equal to  $\mu_0$ , expressed as  $H_0 : \mu = \mu_0$ . In a right-tailed  $t$ -test, the alternative hypothesis  $H_1$  is formulated such that the sample mean is greater than the specified value, given by  $H_1 : \mu > \mu_0$  (Milton and Arnold 1990). The  $t$ -ratio is computed as

$$t = \frac{\mu - \mu_0}{\sigma / \sqrt{n}}, \quad (9)$$

where  $n$  represents the sample size. The next step is to determine whether to reject the  $H_0$ . The  $t$ -ratio follows a  $t_{n-1}$  distribution, where  $n - 1$  is the degrees of freedom. To make a decision, the calculated  $t$ -ratio is compared with critical values from the  $t_{n-1}$  distribution. If the calculated  $t$ -ratio falls into the critical region (extreme tail of the distribution), the null hypothesis is rejected, suggesting that the observed mean is significantly different from the specified value. The critical values are determined based on the desired level of significance (denoted as  $\alpha$ ). If the  $t$ -ratio is beyond the critical values, the null hypothesis is rejected (Milton and Arnold 1990).

The two-sample  $t$ -test is useful for comparing two sets of samples, such as metrics from two ML models, to determine which model is consistently better. When multiple models need comparison, **analysis of variance** (ANOVA) becomes a suitable choice. However, these tests assume normal distribution in the data. When this assumption is inappropriate, nonparametric statistical tests, which require no assumptions about the underlying distribution, can be applied. Examples of nonparametric tests include the **Wilcoxon signed-rank test**, which serves a purpose equivalent to the  $t$ -test, the **Mann-Whitney  $U$  test** for comparing two population medians, and the **Kruskal-Wallis test**, practically a nonparametric version of ANOVA (Milton and Arnold 1990; Moore, McCabe, and Craig 2009).

## 2.4 Preparing the data

In predictive modeling, **data preparation**, also known as **preprocessing**, is a crucial step that typically continues alongside the prediction task. It is not a phase that concludes prior to the commencement of predictions. However, certain **data cleaning** activities, such as rectifying or discarding obviously incorrect values or

variables with uniform values that offer no significant information, can be completed in advance. Common data preparation tasks include data transformation and the handling of missing and/or imbalanced data.

**Data transformation** aims to standardize the scales across various variables, rendering them directly comparable. Most predictive modeling methods require data transformation prior to application. There are numerous approaches to data transformation.

For nominal variables, **one-hot encoding** is a common transformation technique (Rodríguez et al. 2018). This process converts nominal values into binary form. For instance, a nominal variable with three potential values—red, green, and blue—is transformed by creating three binary variables. Each binary variable corresponds to one color, answering the question, “Is the color red/green/blue?”

A prevalent transformation for continuous data is **z-score** standardization, where the transformed value of a variable,  $x_{i,j}$ , is calculated as follows:

$$x_{i,j} = \frac{v_{i,j} - \mu_j}{\sigma_j}, \quad (10)$$

where  $v_{i,j}$  represents the original value,  $\mu_j$  is the mean of all values of the  $j^{\text{th}}$  variable, and  $\sigma_j$  is the standard deviation of the  $j^{\text{th}}$  variable. In predictive modeling, to prevent data leakage, it is crucial to avoid mixing training and testing data during transformation. Transformations should be applied first to the training data, with the derived standardization parameters ( $\mu, \sigma$ ) used for subsequent transformations (Kantardzic 2020). This precaution ensures that the testing data remain independent and unbiased. For binary variables, Gelman and Hill (2006) suggested using  $2\sigma$  instead of  $\sigma$  as the denominator in Eq. 10.

Additional data preparation steps may include eliminating irrelevant variables and erroneous values, generating new variables, addressing missing data, and data reduction (Kuhn and Johnson 2016).

**Principal component analysis** (PCA) is a popular data reduction technique that projects the original variables onto a lower-dimensional space. It achieves this by identifying orthogonal linear combinations of the original variables that capture the majority of the variance. The resulting principal components (PCs) represent the data in a condensed form, ordered by their ability to explain variance—that is, the first PC accounts for the most variance, followed by the second, and so on (Kantardzic 2020; Zaki, Meira Jr, and Meira 2014). This property of PCA is evident in the example provided in Section 3.2, Figure 7, where the variance among observations is noticeably higher along the first PC.

**Correlation metrics** offer insights into relationships between variables in data. They can aid in data preparation, allowing the identification of redundant variables that carry similar information and can be removed. The **Pearson correlation** coefficient,  $\rho \in [-1, 1]$ , quantifies the relationship between two variables,  $x_{:,1}$  and  $x_{:,2}$ , in a data sample. It is defined as follows (Zaki, Meira Jr, and Meira

2014):

$$\rho(\mathbf{x}_{:,1}, \mathbf{x}_{:,2}) = \frac{\sum_{i=1}^N (x_{i,1} - \mu_1)(x_{i,2} - \mu_2)}{\sqrt{\sum_{i=1}^N (x_{i,1} - \mu_1)^2 \sum_{i=1}^N (x_{i,2} - \mu_2)^2}}. \quad (11)$$

Here,  $\mu_1$  and  $\mu_2$  represent the mean values of the respective variables. The continuous correlation coefficient  $\rho$  denotes the strength and direction of a linear relationship—that is, a high positive value signifies a positive correlation, a low value indicates a negative correlation, and  $\rho = 0$  represents independence between variables (Alpaydın 2014).

When conducting multiple statistical tests, such as correlation calculations, employing the **Bonferroni correction** can help address the issue of chance findings indicating significance. For instance, a conventional significance level  $\alpha$  of 0.05 can be adjusted to a Bonferroni-corrected threshold  $\alpha_B$  by dividing  $\alpha$  by the number of tests ( $n$ ) performed, as follows:  $\alpha_B = \alpha/n$ . Subsequently, a significance test is conducted by comparing the obtained p-value against the adjusted threshold  $\alpha_B$  to determine the significance of the result (Alpaydın 2014; Zaki, Meira Jr, and Meira 2014).

**Missing data** is a prevalent challenge in real-world datasets, including those in the health sector. Fragmentation of health services often leads to incomplete data due to the disparate collection and storage practices among various providers, with different variables collected for distinct patient populations (Mirkes et al. 2016). This fragmentation complicates efforts to establish unified global standards.

A typical response to missing data is to exclude observations where it occurs. However, this approach can introduce biases if the underlying **missing data mechanism** is not carefully considered. The following three such mechanisms are recognized (Van Buuren 2012):

- **Missing completely at random (MCAR):** The likelihood that missing data is uniform across all observations, unrelated to the data itself.
- **Missing at random (MAR):** The probability that missing data is consistent within specific groups, dependent on a known attribute.
- **Missing not at random (MNAR):** The probability that missing data varies for unknown reasons.

Occasionally, the absence of data itself can contain information for the modeling process (Kuhn and Johnson 2016). When neither deletion nor direct use of missing observations is viable, **imputation** is a strategy to estimate missing values. The simplest imputation method is to replace missing values with the mean of existing data, which skews the distribution and is generally discouraged (Van Buuren 2012). More sophisticated techniques include regression (refer to Section 3.1),  $k$ -nearest neighbors ( $k$ NN) (refer to Section 3.5), and random forest (RF) (refer to Section 3.4) for missing data (Tang and Ishwaran 2017).

For longitudinal data, **interpolation** techniques can estimate values for missing time points, enabling comparisons between observations with data collected

at different times (Gnauck 2004). Simple interpolation methods include carrying forward the last known observation or obtaining the value from the nearest known neighbor (Gnauck 2004; Kuhn and Johnson 2016; Van Buuren 2012). Linear interpolation involves drawing a straight line between two known observations, providing imputed values for the missing point or points in a straightforward manner (Gnauck 2004). More advanced methods introduce, for example, polynomials to the interpolation functions and can also support multivariate data (Gnauck 2004; Olver 2006).

An alternative to the aforementioned methods is **multiple imputation**, a process involving three key steps (Ginkel et al. 2020). Initially, multiple complete datasets are generated by imputing several values for each missing entry using a statistical model. These datasets are then analyzed independently using standard procedures. Finally, the results from these analyses are combined into a single statistical interpretation (Ginkel et al. 2020). **Multiple imputation using chained equations** (MICE) is a common implementation of this approach (Slade and Naylor 2020; Van Buuren 2012). MICE offers various options for the initial statistical model, including RF (Slade and Naylor 2020). Despite its complexity, multiple imputation provides advantages, such as incorporating the uncertainty of imputed values into the method itself (Van Buuren 2012).

**Imbalanced data** is another common issue. In the case of binary data, this imbalance occurs when the distribution of response variables is skewed, often resulting in an overrepresentation of the majority class and a scarcity of observations for the minority class, which is typically of greater interest (Chawla et al. 2002). Such imbalances can compromise model performance, leading to metrics that are not well-calibrated, such as high specificity paired with low sensitivity (Kuhn and Johnson 2016).

Adjusting the classification threshold is a straightforward method for addressing imbalanced data. The default threshold for binary classification is usually 0.5, meaning an observation is classified as malignant if its probability is 0.5 or higher. Lowering this threshold for the minority class can improve model calibration. Optimal thresholds can be determined through ROC analysis (refer to Section 2.3) (Kuhn and Johnson 2016) or by incorporating domain expertise (Wynants et al. 2019).

For instance, as seen in Figure 1, a classifier's most critical errors are FNs, where malignant cases are incorrectly classified as benign. Some predictive modeling techniques allow users to assign different costs to each type of classification error, such as assigning a higher cost to FNs, which can be beneficial in imbalanced datasets (Kantardzic 2020).

Basic sampling methods, such as **oversampling** (duplicating observations for the minority class) or **undersampling** (removing observations from the majority class), can also be used. However, oversampling can lead to overfitting (refer to Section 2.5) (Kantardzic 2020), while undersampling may result in the loss of valuable information (Kantardzic 2020).

**The synthetic minority oversampling technique** (SMOTE) (Chawla et al. 2002) addresses these issues by combining undersampling of the majority class



with oversampling of the minority class, using the  $k$ NN algorithm (refer to Section 3.5) to generate new synthetic observations. For datasets with mixed variable types, SMOTE-NC (Chawla et al. 2002) and other variations (Rodríguez-Torres, Carrasco-Ochoa, and Martínez-Trinidad 2021) have been developed to effectively counteract class imbalance.

## 2.5 Model selection and assessment

In predictive modeling, accurately estimating a model’s predictive capability with independent **testing data** is crucial (Bishop 2006; Hastie, Tibshirani, and Friedman 2009). Validation serves a dual purpose; it not only gauges a model’s performance but also aids in selecting the most suitable model from a set of candidates (Kohavi 1995). Moreover, validation reveals the model’s **generalization** ability—that is, its anticipated efficacy with novel data (Bishop 2006; Hastie, Tibshirani, and Friedman 2009). In contrast, explanatory modeling does not involve validation with independent testing data, which may lead to an overestimation of its predictive accuracy for new observations (Sainani 2014).

**Model selection** involves evaluating various candidate models to identify the most effective one (Hastie, Tibshirani, and Friedman 2009). Additionally, **model assessment** refers to estimating the performance of the optimal model on unseen data (Hastie, Tibshirani, and Friedman 2009). In the context of health applications, creating a universally generalizing model is often impractical due to the multitude of patterns and local nuances influencing model development. Nonetheless, such models can still hold clinical value. In such instances, models may require retraining to adapt to local conditions (Futoma et al. 2020). Furthermore, a model demonstrating strong results may necessitate ongoing updates to maintain its predictive power (Ghassemi et al. 2020). During model validation, performance metrics represent a balance between competing objectives, such as sensitivity and specificity.

Figure 3 depicts the relationship between model complexity and classification performance, highlighting the critical juncture at which further training does not improve generalization performance. When the MCC for the training data reaches 1, the model becomes **overfit** to the training data without enhancing its generalization capability. Instead, with visual inspection, there seems to be an increase in variance and a subtle downward trend in generalization performance, as reflected by the mean validation MCC.

Conversely, **underfitting** occurs in Figure 3 when the tree’s maximum depth is set to 1 or 2. Another aspect of escalating model complexity is the initial high bias and low variance of simple models. As complexity grows, bias decreases, but variance increases (Hastie, Tibshirani, and Friedman 2009). To prevent overfitting, **regularization** techniques are employed. These methods impose penalties on model complexity, resulting in simpler models with enhanced generalization capabilities (Kuhn and Johnson 2016).

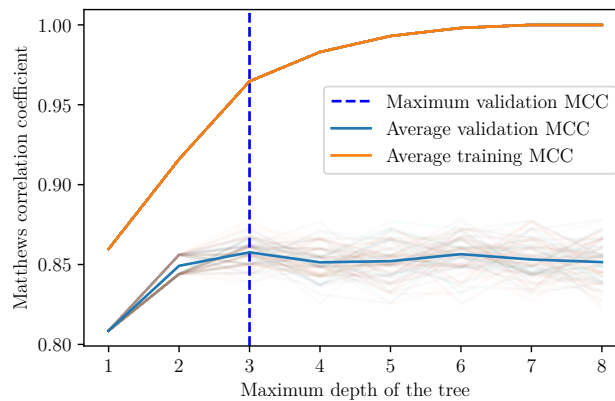


FIGURE 3 As model complexity increases, a decline in the model’s generalization performance is often observed (Hastie, Tibshirani, and Friedman 2009). In this instance, a decision tree classifier (refer to Section 3.3) was employed to discern the two classes within the breast cancer data (Wolberg et al. 1995). This example, adapted from the documentation by Pedregosa et al. (2011), showcases the relationship between model complexity and performance. The primary indicators, average validation Matthews correlation coefficient (MCC) and average training MCC, represent the mean values from one hundred iterations of 10-fold cross-validation. These iterations are visually represented as faint lines surrounding the average validation curve. The critical point occurs when the tree’s maximum depth is set to three—here, the validation MCC peaks, indicating optimal generalization performance. Beyond this depth, as complexity increases, the validation MCC no longer increases. Conversely, the training MCC continues to ascend, ultimately reaching the maximum value of one, signifying perfect training fit.

In a best case scenario, there are many data that can be utilized when training and validating the models. Then, a portion of data can be taken out as a final testing set and kept aside until the very end of the modeling process (Hastie, Tibshirani, and Friedman 2009). However, this is often not possible, and the overall data might have to be used in making use of the metrics gathered during **cross-validation**.

### 2.5.1 Cross-validation and bootstrap

Cross-validation (CV) is a technique used to assess a model’s predictive performance on unseen data. By conducting multiple validation iterations, CV provides an average performance estimate that reflects the model’s ability to generalize. In  $k$ -fold CV, depicted in Figure 4, the dataset is randomly partitioned into  $k$  equally sized subsets, or folds, ensuring each observation is exclusive to a single fold (Kohavi 1995). During the  $k$ -fold CV process, each fold sequentially serves as the testing set, while the remaining folds constitute the training set (Hastie, Tibshirani, and Friedman 2009; Kohavi 1995). This cycle is repeated  $k$  times, allowing each fold to be used for testing once.

To ensure unbiased performance estimates, the testing data must not influ-

Fold	Data split			
1	Testing	Training		
2	Training	Testing	Training	
3	Training		Testing	Training
4	Training		Testing	Training
5	Training			Testing

FIGURE 4 Data split in a 5-fold cross-validation.

ence the training phase. Within the training data of a fold, there can be various ways of selecting the model whose performance is estimated with the fold’s testing data.

For instance, in the RF method (refer to Section 3.4), not all observations are used during training. The unused observations can act as **validation data**, offering an unbiased estimate of the model’s generalization capability (Breiman 2001a). An **inner** or **nested**  $k$ -fold CV can also be implemented within each fold, further dividing the training data into sub-training and validation sets. The benefit of this rather complex arrangement ensures that the performance estimate, derived from multiple runs, is not influenced by the testing data (Varma and Simon 2006). Hyperparameter optimization, discussed in Section 2.6, is often conducted at this stage.

However, a recent empirical study by Wainer and Cawley (2021) suggested that nested CV may be excessive for most practical applications. A simpler, non-nested CV approach, where model selection is based on the testing data, may suffice for selecting optimal hyperparameters and determining the best predictive method.

Performance measures are recorded for each fold. The estimated CV performance measure,  $E[\theta]$ , is the mean of all  $k$  fold estimates, as follows (Zaki, Meira Jr, and Meira 2014):

$$E[\theta] = \frac{1}{k} \sum_{i=1}^k \theta_i, \quad (12)$$

where  $\theta$  is the selected performance metric (e.g., MCC).

The stability of the estimate across folds is also important; high variance indicates unreliability, while low variance denotes a more trustworthy estimate (Kohavi 1995). The variance estimate,  $\sigma_\theta^2$ , is calculated as follows (Zaki, Meira Jr, and Meira 2014):

$$\sigma_\theta^2 = \frac{1}{k} \sum_{i=1}^k (\theta_i - E[\theta])^2. \quad (13)$$

Commonly,  $k$  is set to 5 or 10, balancing bias and variance effectively (Hastie, Tibshirani, and Friedman 2009). Some studies have also highlighted the importance of repeating  $k$ -fold CV multiple times (Bouckaert 2003; J.-H. Kim 2009), and

specifically in nested CV (Krstajic et al. 2014), to get more reliable estimates. **Stratification** during CV ensures a consistent distribution of output variables across folds, reducing estimate variance and bias, thereby enhancing reliability (Kohavi 1995).

Leave-one-out CV represents a special case of  $k$ -fold CV where  $k$  equals  $N$  of the dataset. In it, each test set includes a single data observation, with the remainder forming the training set. Although leave-one-out CV generally exhibits low bias, its high variance can render the estimates less reliable (Hastie, Tibshirani, and Friedman 2009; Kohavi 1995). The choice of the CV method depends on the nature of the problem. For instance, leave-one-out CV is commonly employed when dealing with small datasets (Wong 2015).

The validation approaches described above are categorized as **internal**, indicating that the test sets originate from a singular dataset. For a more robust evaluation of model performance, **external** validation is recommended, utilizing an entirely distinct dataset, such as a different cohort (Van Calster et al. 2019).

Beyond CV, the bootstrap method also serves to estimate performance. It creates new datasets by sampling with replacements from the training data, allowing for the possibility of multiple inclusions of a single observation within a bootstrap sample (Hastie, Tibshirani, and Friedman 2009). While bootstrap does not meet the independent validation criteria for predictive modeling, it finds use in methodologies like RF (refer to Section 3.4), leveraging out-of-bag observations to assess model performance.

## 2.6 Hyperparameter optimization

Hyperparameters, such as the maximum depth of a decision tree (refer to Section 3.3), are crucial in shaping a model's structure. These parameters must be established by the user before initiating the training of an ML model (Bischl et al. 2023; Yang and Shami 2020). Prior to **hyperparameter optimization**, one must select a target metric (e.g., one of the performance measures discussed in Section 2.3) and the hyperparameters to be optimized.

Suppose we have a dataset  $\mathbf{D} = \{\mathbf{x}_i, y_i\}_{i=1}^N$ , where  $\mathbf{x}_i$  is an input vector,  $y_i$  is the corresponding scalar output in the set of  $D$ , and  $N$  represents the total count of observations. The challenge of optimizing the hyperparameter configuration  $\lambda$ , which belongs to the comprehensive hyperparameter space  $\Lambda$ , particularly in a classification context where an increase in the target metric signifies an improved classifier, can be formulated as (Alpaydm 2014)

$$\lambda^* = \arg \max_{\lambda \in \Lambda} E_{val}(\lambda | \mathbf{D}), \quad (14)$$

where  $\lambda^*$  denotes the optimal hyperparameter configuration and  $E_{val}(\cdot)$  is a function that yields the desired classification metric for  $\lambda$ , given an independent validation dataset  $D$ . Essentially, during optimization, we evaluate various potential hyperparameter configurations to determine which  $\lambda$  delivers the optimal

result for our target metric (Alpaydm 2014). CV can be utilized to estimate generalization performance, where the ideal hyperparameter configuration maximizes, on average, the validation metric (Bergstra and Bengio 2012). Common hyperparameter optimization techniques include grid search, random search, and Bayesian optimization. All these techniques can be utilized in conjunction with CV.

**Grid search** involves predefining all relevant hyperparameters and their precise values before commencing optimization. This method exhaustively explores the grid of hyperparameter configurations during the optimization process. Grid search is straightforward to implement; however, as the number of hyperparameter configurations grows, so too does the computational time (Bischl et al. 2023; Lorenzo et al. 2017; Yang and Shami 2020).

**Random search**, a variant of grid search, does not examine every possible hyperparameter combination but instead evaluates a predefined number of random configurations (Bischl et al. 2023; Yang and Shami 2020). This approach is more practical when the configuration search space is extensive.

**Bayesian optimization** (BO) represents a more sophisticated method for navigating vast hyperparameter spaces. Unlike the aforementioned techniques, BO uses the outcomes of the search to inform future configurations, concentrating time and effort on areas of the search space more likely to yield improvements for the chosen metric (Yang and Shami 2020). Typically framed as a minimization problem, the **cost** of a configuration is determined by inverting the desired metric. BO is particularly beneficial for optimizing resource-intensive black-box functions, where each hyperparameter configuration evaluation incurs significant expense (Shahriari et al. 2016).

BO encompasses two primary components. The first is the **surrogate function**, a probabilistic model that serves as a simplified representation, or approximation, of the complex true objective function. Within Bayesian terminology, this surrogate function is referred to as the **prior**. The second component involves the strategic exploration of the surrogate function through an **acquisition function**. The acquisition function's role is to determine the next hyperparameter configuration to be investigated (Shahriari et al. 2016).

A common surrogate function utilized in BO is the **Gaussian process** (GP). GP presumes a multivariate normal distribution, so it is mainly used for optimizing continuous variables (Rasmussen and Williams 2006). Initially, GP selects a set of random functions as priors. Essentially, GP's task is to find out which of these random function realizations could have produced the observed values—that is, the chosen target metrics (Gramacy 2020). With each exploration of a new configuration, the prior is updated, evolving into the Bayesian **posterior** distribution, which encapsulates the optimizer's updated understanding of the objective function based on the data observed (Shahriari et al. 2016). An illustration of GP within the context of BO is depicted in Figure 5.

GP is characterized by a mean vector,  $\mu$ , and a covariance or kernel function,  $k(\lambda, \lambda')$ . GP generates random outputs,  $Y$ , that adhere to a normal distribution,

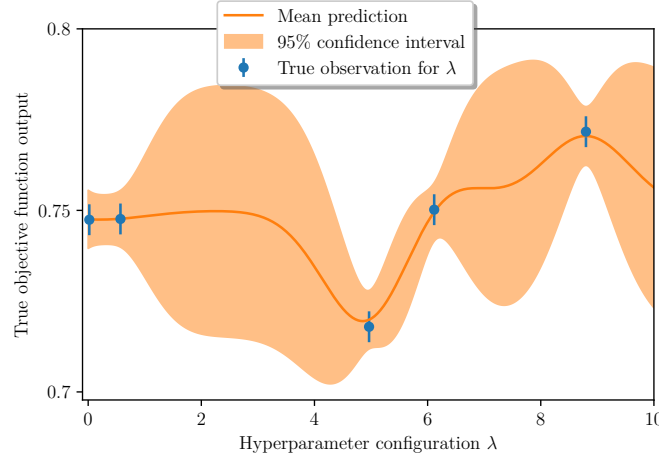


FIGURE 5 A simplified example of a Gaussian process in Bayesian optimization (BO), adapted from Pedregosa et al. (2011). The x-axis symbolizes the array of potential hyperparameter configurations,  $\lambda$ , for evaluating the true objective function. Proximity on the x-axis indicates similarity between hyperparameter configurations, while greater distances suggest dissimilarity. The y-axis denotes the outcome of the objective function—that is, the chosen performance metric. Assuming the metric is a Matthews correlation coefficient (refer to Section 2.3), a higher metric value signifies a better classifier. So far, five hyperparameter configurations have been assessed, marked as true observations for  $\lambda$ . The acquisition function in BO is tasked with selecting the next  $\lambda$  for exploration. In this scenario, a promising area for exploration might be around the x-axis values of 7–8, where the upper confidence interval exceeds the highest observed output, indicating some potential for enhancement.

expressed as

$$Y \sim GP(\boldsymbol{\mu}, k(\boldsymbol{\lambda}, \boldsymbol{\lambda}')), \quad (15)$$

where typically  $\boldsymbol{\mu} = 0$ , signifying that the data are centered. The kernel function  $k$  accepts two inputs, each representing a hyperparameter configuration  $\boldsymbol{\lambda}$  with  $N$  hyperparameters. This function yields an  $N \times N$  covariance matrix that describes the interrelations among each pair of hyperparameters (Gramacy 2020). The kernel function ensures that similar hyperparameter configurations produce closely related outputs. There exists a variety of kernel functions; for example, MATLAB employs the ARD Matérn 5/2 kernel (MathWorks 2023), which is favored for practical optimization tasks due to its generation of functions that are not unrealistically smooth (Snoek, Larochelle, and Adams 2012). Additionally, it is presumed that the observations include Gaussian noise (MathWorks 2023).

Among the most popular acquisition functions is the **expected improvement** (EI) (Feurer and Hutter 2019; Jones, Schonlau, and Welch 1998). The EI for a hyperparameter configuration  $\boldsymbol{\lambda}$  is expressed as

$$E[I(\boldsymbol{\lambda})] = E[\max(f^* - Y, 0)], \quad (16)$$

where  $f^*$  represents the optimal observed value of the objective function thus far. An enhancement in the outcome (reducing cost) is anticipated when  $f^* - Y > 0$ ;

otherwise, no improvement is expected upon exploring the hyperparameter configuration with the true objective function. Consequently, the true objective function is assessed only for those configurations projected to yield enhancements for the chosen metric (Feurer and Hutter 2019; Jones, Schonlau, and Welch 1998; Shahriari et al. 2016).

Furthermore, enhancements to the original EI acquisition function have been proposed. The default configuration of MATLAB incorporates two such improvements—namely, **per second** and **plus** (MathWorks 2023). The per-second modification considers the computational cost of evaluating a configuration, with the aim of minimizing computational time. This is achieved by managing a separate GP for the evaluation time while optimizing the performance metric (Snoek, Larochelle, and Adams 2012). The plus modification, on the other hand, is designed to prevent excessive exploitation of a particular search area. It actively seeks a superior global solution to the minimization problem, avoiding entrapment in the local search area (Bull 2011; MathWorks 2023).

## 2.7 Explaining the predictions

Enhancing the predictive capabilities of models is often insufficient, particularly in the health sector, where ethical and legal considerations must be factored into the clinical application of predictive models. Even when dealing with a black box model, obtaining explanations and justifications for its predictions can be beneficial and, sometimes, legally mandated (Payrovnaziri et al. 2020; Vellido 2020). For instance, the **General Data Protection Regulation** (GDPR) requires data controllers to provide information about the logic behind automated decision-making processes and offer justifications for the resulting outcomes (Hamon et al. 2022). Critics argue that reliance on black box models has led to significant societal issues, impacting areas such as health, freedom, racial bias, and safety (Rudin 2019; Rudin et al. 2022).

Certain predictive modeling methods offer greater interpretability, with explanations often embedded within the models themselves. For example, logistic regression (refer to Section 3.1) utilizes coefficient terms that are interpretable with basic methodological knowledge. Decision trees (see Section 3.3) are even more accessible, understandable by novices when classification rules are visually presented. However, as model complexity increases, so too does the challenge of interpreting predictions (Ahmad, Eckert, and Teredesai 2018; Vellido 2020). For complex models, mechanisms for explaining predictions may exist, such as the list of contributing predictors produced by RF (see Section 3.4).

An ML model is deemed interpretable if it provides explanations or rationales alongside predictions or recommendations (Ahmad, Eckert, and Teredesai 2018). Interpretability is highly domain-specific, with varying constraints across different domains (Rudin et al. 2022). In interpretable ML, these constraints are fundamental to the optimization problem. For instance, a decision tree with

fewer leaves is simpler to interpret, so a penalty for increasing the number of leaves can be incorporated during tree construction (Rudin et al. 2022).

An alternative approach is the post-hoc explanation of black box model predictions, known as **explainable AI** (Rudin et al. 2022). Methods such as **local interpretable model-agnostic explanations** (Ribeiro, Singh, and Guestrin 2016) and **Shapley additive explanations** (Lundberg and Lee 2017) can elucidate predictions from any model.

Recent studies have highlighted that simply offering explanations for predictions may not significantly enhance human users' decision-making processes, since individuals often underutilize available explainability tools (Miller 2019, 2023). Drawing on extensive research from disciplines such as psychology and philosophy could aid in addressing trust issues related to AI. For instance, a proposed framework known as **evaluative AI** suggests that decision support tools should present evidence both supporting and opposing human decisions, rather than merely attaching explanations to predictions (Miller 2023).

Additionally, contemporary research has delved into expressing a system's confidence in its predictions to users (Waa et al. 2020), harmonizing the somewhat incoherent terminology within the field (Graziani et al. 2022), and identifying both challenges and prospects specific to ML applications in the clinical medical domain (Antoniadi et al. 2021). These efforts contribute to the ongoing discourse on enhancing the interpretability and trustworthiness of AI systems in critical sectors.



## 3 PREDICTION METHODS

This chapter presents the predictive modeling methods employed in this thesis. Linear and logistic regression are explored in Section 3.1, followed by support vector machines in Section 3.2. Decision trees and RFs are examined in Sections 3.3 and 3.4, respectively. The chapter concludes with a brief introduction to two additional methods, in Section 3.5.

### 3.1 Linear and logistic regression

**Linear regression** models are favored for their simplicity and interpretability. Such models' influence on the output is discernible through their regression coefficients. Moreover, linear regression can be utilized to preprocess input values, even when alternative methods are applied for the actual prediction task (Hastie, Tibshirani, and Friedman 2009).

Consider an input vector,  $\mathbf{x} = (x_1, x_2, \dots, x_p)$ , with  $p$  variables. The predicted scalar output  $\hat{y}_i$  is computed using a linear regression model, which is formulated as follows (Hastie, Tibshirani, and Friedman 2009):

$$\hat{y}_i = \beta_0 + \sum_{j=1}^p x_{i,j} \beta_j = \boldsymbol{\beta}^T \mathbf{x}_i. \quad (17)$$

$\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_p)$  are the regression coefficients established during the model's training phase,  $\beta_0$  being the intercept term, and  $\mathbf{x}_i = (1, x_{i,1}, x_{i,2}, \dots, x_{i,p})$  is the  $i^{\text{th}}$  input vector augmented by value 1 in the first element for the intercept term.

The **least squares** method is commonly employed to estimate these  $\boldsymbol{\beta}$  coefficients, aiming to minimize the sum of squared errors (SSE), as follows (Hastie, Tibshirani, and Friedman 2009; Kuhn and Johnson 2016; Tan et al. 2013):

$$\text{SSE}(\boldsymbol{\beta}) = \sum_{i=1}^N (y_i - \hat{y}_i)^2, \quad (18)$$

where  $y_i$  represents the true response variable for a single observation  $\mathbf{x}_i$  (Hastie, Tibshirani, and Friedman 2009). By minimizing the SSE, we achieve the most accurate linear approximation of the training data.

**Logistic regression** (LR) is a widely-used classification method that estimates the probability  $p$  of an observation belonging to a specific class. For binary classification, the probability of class 0 is given by (Hastie, Tibshirani, and Friedman 2009)

$$p(y_i = 0|\mathbf{x}_i) = \frac{1}{1 + e^{-(\boldsymbol{\beta}^T \mathbf{x}_i)}}. \quad (19)$$

To create models suitable for multi-variable datasets and prevent overfitting, regularization techniques are employed (Koh, S.-J. Kim, and Boyd 2007).  $L_1$  regularization constrains coefficient sizes by adding a penalty equal to their absolute values' sum, adjustable via the  $\lambda$  parameter.  $L_1$ -regularized LR, also known as lasso, can be expressed as a maximization problem, as follows (Hastie, Tibshirani, and Friedman 2009):

$$\max_{\beta_0, \boldsymbol{\beta}} \left\{ \sum_{i=1}^N \left[ y_i \left( \beta_0 + \sum_{j=1}^p x_{i,j} \beta_j \right) - \log \left( 1 + \exp \left( \beta_0 + \sum_{j=1}^p x_{i,j} \beta_j \right) \right) \right] - \lambda \sum_{j=1}^p |\beta_j|^r \right\}, \quad (20)$$

where  $\lambda \sum_{j=1}^p |\beta_j|^r$  is the regularization term with  $r = 1$ . This approach often reduces many coefficients to zero, acting as a variable selection method (Koh, S.-J. Kim, and Boyd 2007). When the variables are strong but correlated with each other, this penalty is described as "somewhat indifferent" to the choice of variables (Hastie, Tibshirani, and Friedman 2009).

Conversely,  $L_2$  regularization, known as ridge regression, is applied when  $r = 2$ . Unlike  $L_1$ , it does not reduce any of the coefficients to zero, retaining all variables in the model (Koh, S.-J. Kim, and Boyd 2007). This penalty tends to shrink coefficients of correlated variables towards each other (Hastie, Tibshirani, and Friedman 2009).

## 3.2 Support vector machine

**Support vector machine (SVM)** is a widely-used method that, for a binary classification problem, aims to create a hyperplane that effectively separates the two classes. SVM for classification is a **support vector classifier (SVC)**. For datasets with only two variables, the separating hyperplane can be visualized as a line demarcating the classes, exemplified by the solid line between two groups of observations in Figure 6. Predictions for new data points are determined based on their placement relative to this hyperplane.

While more complex to visualize, the fundamental concept of a separating hyperplane extends to higher dimensions as well (Noble 2006). Formally, the hyperplane  $f(\mathbf{x})$  is defined by the following equation (Hastie, Tibshirani, and Friedman 2009):

$$f(\mathbf{x}) = \mathbf{x}^T \boldsymbol{\beta} + \beta_0 = 0, \quad (21)$$

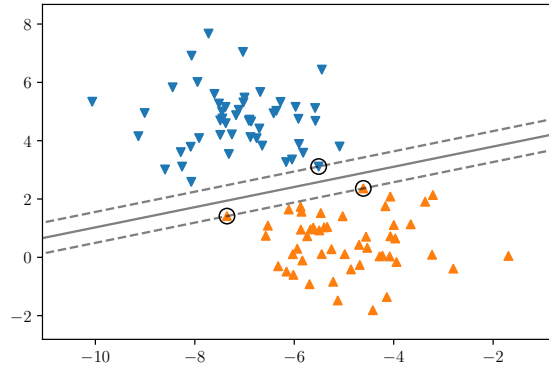


FIGURE 6 An example of a linear support vector classifier (Pedregosa et al. 2011) where the two classes are linearly separable. The separating hyperplane, drawn in solid line, is also the maximum margin hyperplane. Support vectors are indicated as circled observations.

where  $\mathbf{x}$  represents an observation from the training data,  $\boldsymbol{\beta}$  is a unit vector of length 1 that holds the weights for the  $p$  variables in  $\mathbf{x}$ , and  $\beta_0$  is a scalar bias term.

The best separating hyperplane, known as the **maximum margin hyperplane**, not only separates the classes but also maximizes the distance between them (Noble 2006). The support vectors lie on the two hyperplanes that flank the maximum margin hyperplane. Given class labels  $y_i \in \{-1, 1\}$ , the task of identifying this hyperplane can be defined as a minimization problem, as follows (Hastie, Tibshirani, and Friedman 2009):

$$\min_{\boldsymbol{\beta}, \beta_0} \|\boldsymbol{\beta}\| \quad \text{subject to} \quad y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq 1, \quad i = 1, \dots, N. \quad (22)$$

In scenarios where perfect class separation is unattainable, SVC allows for a **soft margin**. This margin acts as a threshold, permitting a certain degree of misclassification during model construction. Incorporating this soft margin modifies the minimization problem, as follows (Hastie, Tibshirani, and Friedman 2009):

$$\begin{aligned} \min_{\boldsymbol{\beta}, \beta_0} \quad & \frac{1}{2} \|\boldsymbol{\beta}\|^2 + C \sum_{i=1}^N \zeta_i \\ \text{subject to} \quad & \zeta_i \geq 0, y_i(\mathbf{x}_i^T \boldsymbol{\beta} + \beta_0) \geq 1 - \zeta_i \quad \forall i. \end{aligned} \quad (23)$$

This revised equation introduces a cost or regularization parameter  $C$ , which quantifies the penalty for constraint violations, while  $\zeta_i$  represent the distances of training data points that breach these constraints (Kantardzic 2020). The soft margin tolerates some training data misclassifications. Typically, a higher  $C$  value compels the SVM to minimize training misclassifications, although at the expense of the model's generalization capabilities for new data (Hastie, Tibshirani, and Friedman 2009; Kantardzic 2020).

When dealing with linearly non-separable classes, where a straightforward line or hyperplane fails to separate the data, a kernel function can be utilized to

map the data into a higher-dimensional space. This function makes SVM nonlinear (Kantardzic 2020; Noble 2006). Introducing a kernel function and reformulating the optimization challenge as a Lagrangian dual problem involves minimizing the Lagrangian multipliers  $\alpha$ , as follows (Hastie, Tibshirani, and Friedman 2009; Kantardzic 2020):

$$\begin{aligned} \min_{\alpha} \quad & \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j k(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} \quad & \begin{cases} \sum_{i=1}^N y_i \alpha_i = 0 \\ 0 \leq \alpha_i \leq C, i = 1, \dots, N, \end{cases} \end{aligned} \quad (24)$$

where  $k(\cdot)$  denotes the kernel function. For instance, a polynomial kernel of degree  $d$  is expressed as  $k(\mathbf{x}, \mathbf{x}') = (1 + \langle \mathbf{x}, \mathbf{x}' \rangle)^d$  (Hastie, Tibshirani, and Friedman 2009). Figure 7 illustrates an SVC employing a polynomial kernel for nonlinearly separable data.

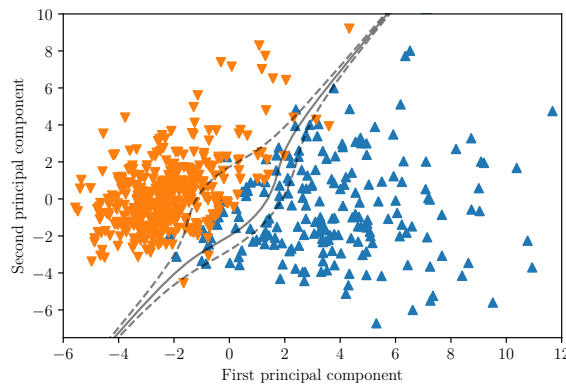


FIGURE 7 A third-degree polynomial kernel support vector classifier (SVC) (Pedregosa et al. 2011) was used to classify the breast cancer data (Wolberg et al. 1995). The data were standardized before principal component analysis was performed (refer to Section 2.4). Only the first two principal components were utilized in the SVC. The soft margin allows some observations to fall behind the wrong sides of the separating hyperplane.

SVM can also be applied for multi-class problems, by forming the classification problem as solving multiple two-class problems, or for predicting continuous responses (Hastie, Tibshirani, and Friedman 2009).

### 3.3 Decision trees

**Decision trees**, renowned for their intuitive and visual nature, establish a set of rules to classify data, grouping observations sharing the same class within each leaf node (Kingsford and Salzberg 2008; Quinlan 1986). While decision trees accommodate both classification and regression, our primary focus remains on classification. An illustrative minimal decision tree is depicted in Figure 8.

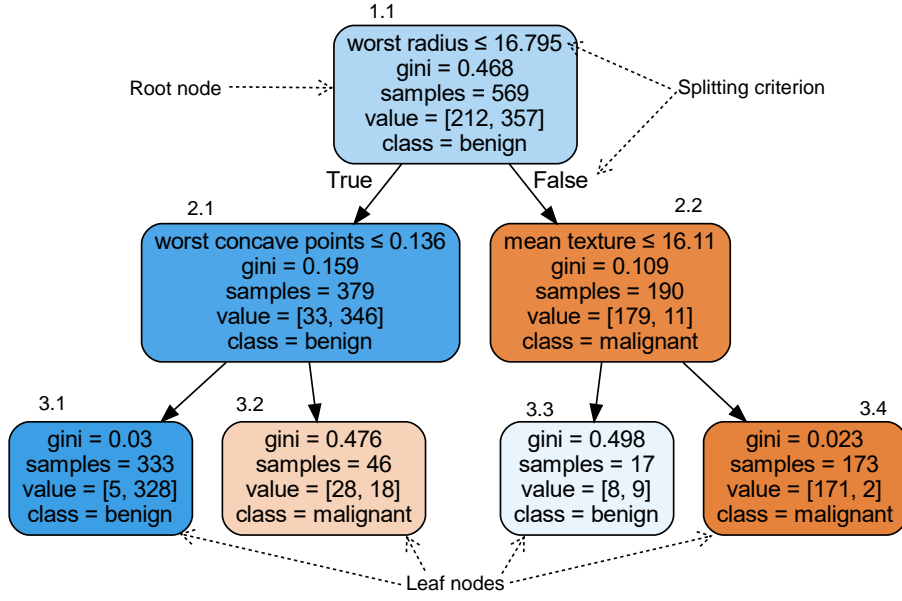


FIGURE 8 An example of a minimal decision tree (maximum depth defined as 2) (Pedregosa et al. 2011) classifying breast cancer data (Wolberg et al. 1995) observations using three variables in data (**worst radius**, **worst concave points**, and **mean texture**). **Root node**, **leaf nodes**, and **splitting criteria** are marked in the figure. Additionally, node 2.1 is an example of a **child node** of 1.1, while 2.2 is a **parent node** of 3.3 and 3.4.

Hunt's algorithm serves as a foundational framework for describing a generic decision tree for classification, forming the basis for many tree implementations (Tan et al. 2013). Let  $D_s = \{x_i, y_i\}_{i=1}^N$  represent the training data in the tree node  $s$ . The iterative process involves the following steps for every child node until further changes are impossible:

1. If all observations in  $D_s$  belong to same class  $y_s$ , node  $s$  becomes a leaf node labeled as class  $y_s$ ;
2. Else, if  $D_s$  contains observations from multiple classes, a **splitting criterion** is applied to divide the observations into smaller subsets, thereby creating child nodes.

The essence of the splitting criterion in the second step is to measure the quality of a split, offering several options. Common criteria include Gini impurity and entropy. Gini impurity  $G$  for data  $D$  is given by (Tan et al. 2013)

$$G(D) = 1 - \sum_{k=1}^K p_k^2, \quad (25)$$

where  $K$  represents the number of classes and  $p_k$  denotes the probability of observing class  $k$ . Additionally, entropy  $H$  for data  $D$  is defined as (Tan et al. 2013)

$$H(D) = - \sum_{k=1}^K p_k \log_2 p_k. \quad (26)$$

Gini impurity or entropy serves as the basis for computing the **information gain**  $\Delta$ , a metric used to assess the quality of a split, as follows (Tan et al. 2013):

$$\Delta = I(\text{parent}) - \sum_{j=1}^k \frac{N(v_j)}{N} I(v_j), \quad (27)$$

where  $I(\text{parent})$  represents the criterion (such as  $G$  or  $H$ ) of the parent node,  $k$  signifies the number of divisions based on the condition (e.g.,  $k = 2$  for a binary tree),  $N(v_j)$  denotes the observations in child node  $j$ ,  $N$  is the total observations in the parent node, and  $I(v_j)$  is the criterion of child node  $j$ . Higher  $\Delta$  values indicate a more favorable split.

In decision trees, overfitting is commonly addressed using two methods: **pre-pruning** and **post-pruning**. Pre-pruning involves halting tree growth before all training data observations are fitted, often by restricting the tree depth or the number of leaf nodes or specifying the minimum observations required for splitting. On the other hand, post-pruning entails growing the entire tree and later removing or combining nodes for overly specific cases (Tan et al. 2013).

Several decision tree algorithms exist, with ID3, CART, and C4.5 being widely used (Hastie, Tibshirani, and Friedman 2009). These individual decision trees form a foundational framework for more sophisticated ensemble methods such as an RF.

### 3.4 Random forest

RF is an ensemble learning method that constructs a collection of uncorrelated decision trees (Breiman 2001a). Utilizing decision trees and **bootstrap aggregation (bagging)**, RF aims to enhance predictions by building a forest of trees that work independently. The forest operates as an ensemble, leveraging the collective wisdom of multiple trees, which helps balance any inaccuracies in individual trees (Breiman 2001a; Hastie, Tibshirani, and Friedman 2009). An illustrative depiction of a small random forest consisting of three trees is presented in Figure 9.

The RF algorithm initializes by creating bootstrap samples with replacement from the training data. For each leaf node in a tree, the following steps are iteratively performed (Breiman 2001a):

1. Randomly select  $m$  variables from a pool of  $p$  potential input variables;
2. Determine the optimal splitting point among the selected  $m$  variables;
3. Split the node into two child nodes.

These steps continue until the minimum node size is reached (Hastie, Tibshirani, and Friedman 2009).

The process described above, starting from creating the bootstrap sample, iterates until the forest consists of the specified number of trees, denoted as  $T$ . The

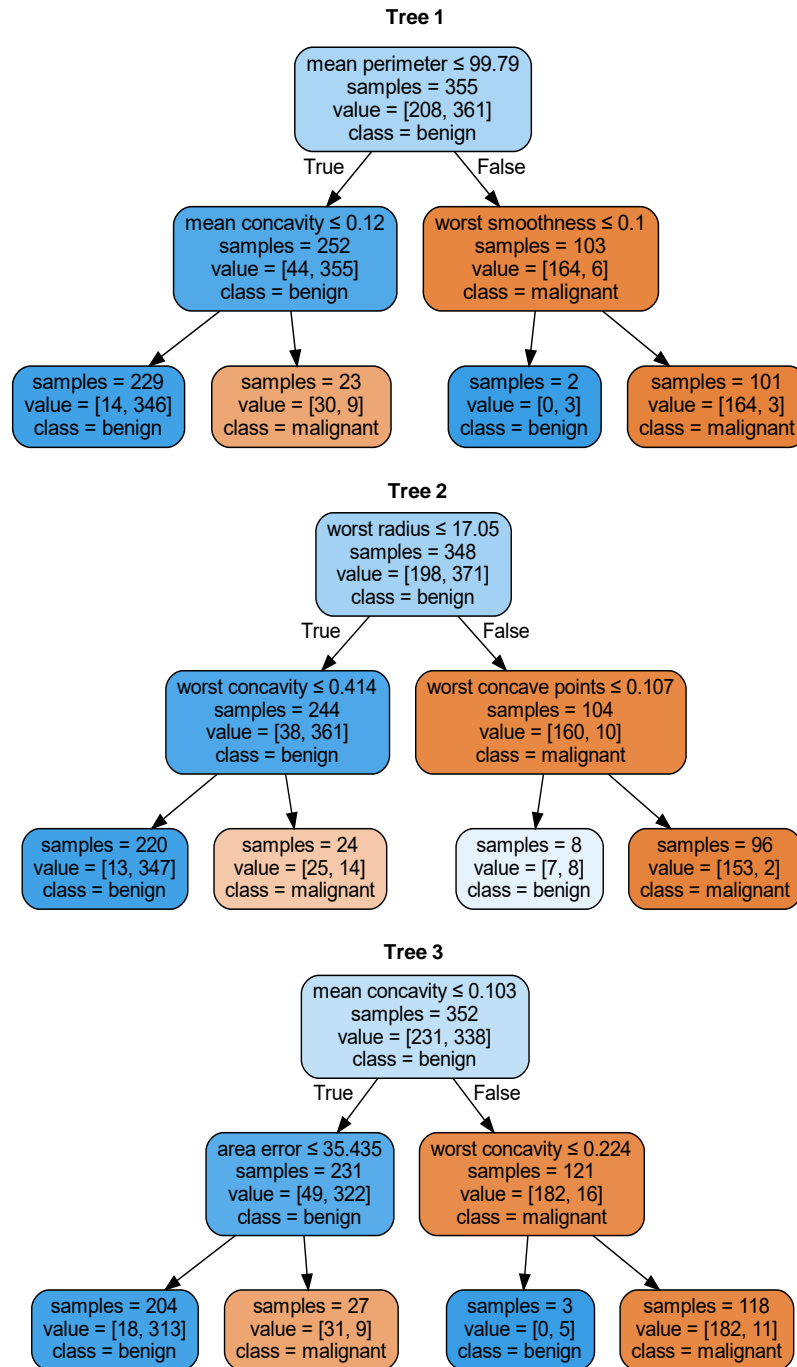


FIGURE 9 A minimal random forest example (with a maximum depth of two per tree and a total of three trees) (Pedregosa et al. 2011) classifying breast cancer data (Wolberg et al. 1995). Each of the three decision trees within the forest is distinct, and seven out of nine variables are used uniquely across the forest.

algorithm's output is the ensemble of trees, constituting the forest. This forest is then utilized for predictions. In regression tasks, each tree predicts an observation's output, and the forest's output is the average of all predictions made by the trees. In classification tasks, the most frequent class among the trees determines the final prediction of the RF (Hastie, Tibshirani, and Friedman 2009). The combination of creating bootstrap samples (bagging) and random variable selection

consistently improves prediction accuracy (Breiman 2001a).

A key advantage of RF lies in requiring tuning for relatively few hyperparameters, and some recommended default values are provided by the author of RF (Breiman 2001a; Hastie, Tibshirani, and Friedman 2009). For the number of randomly chosen variables,  $m$ , the default values are  $m = \lfloor \sqrt{p} \rfloor$  for classification and  $m = \lfloor p/3 \rfloor$  for regression (Hastie, Tibshirani, and Friedman 2009).

One crucial setting in RF is the number of trees within the forest,  $T$ . Extensive studies (Probst and Boulesteix 2018) have discouraged considering the number of trees as a tunable parameter. Generally, more trees lead to better performance, although increased computational cost is the primary drawback. Overfitting is less concerning in RF (Breiman 2001a; Hastie, Tibshirani, and Friedman 2009). However, Probst and Boulesteix (2018) demonstrated that the most significant performance enhancement in classification is observed within the first 100 trees, with subsequent trees offering limited gains in performance.

Empirical evidence suggests that RF hyperparameters are less adjustable compared to some other algorithms (Probst, Boulesteix, and Bischl 2019; Probst, Wright, and Boulesteix 2019). Optimizing these hyperparameters showed only a marginal 0.01 increase in the average AUC value when compared to default configurations. However, this slight increase can be significant in specific cases. They highlighted key RF hyperparameters, including **sample size**, whether to sample **with or without replacement**, **node size**, and **splitting criterion**.

RF can also be utilized for variable importance estimation, helping model interpretation, and aiding in variable selection (Díaz-Urriarte and Alvarez de Andrés 2006; Genuer, Poggi, and Tuleau-Malot 2010). Assessing variable importance in RF is often facilitated through the **out-of-bag** (OOB) error estimation, referring to observations not used in the bootstrap sample.

Let  $\text{OOB}_t$  represent the bootstrap sample for tree  $t$  and  $\text{OOBerr}_t$  denote the tree's error.  $\text{OOBerr}_t$  is calculated by utilizing the tree to predict the classes or values for the OOB observations. For classification, the error corresponds to the misclassification rate, while in regression, it signifies the mean square error.

The **permutation importance** (PI) or **mean decrease in accuracy** for the  $j^{\text{th}}$  variable in the dataset is computed as (Breiman 2001a; Genuer, Poggi, and Tuleau-Malot 2010; Louppe et al. 2013)

$$\text{PI}_j = \frac{1}{T} \sum_t (\widetilde{\text{OOBerr}}_{t,j} - \text{OOBerr}_t), \quad (28)$$

where  $\widetilde{\text{OOBerr}}_{t,j}$  represents the tree's error obtained when the values of the  $j^{\text{th}}$  variable are randomly permuted (Genuer, Poggi, and Tuleau-Malot 2010) and  $T$  is the number of trees within the forest. In essence, this process involves randomly shuffling the values of the  $j^{\text{th}}$  variable among OOB observations, predicting with a single tree, recording the error rate, and subtracting the original OOB error. The average of these differences across all  $T$  trees in the forest yields the final importance value for the  $j^{\text{th}}$  variable in the dataset. An importance value around zero indicates lack of predictive power.



Although originating from RF, PI is not limited to this model; it can be applied across various methods. However, in the absence of OOB observations in most algorithms, separate validation/test data become essential (Molnar et al. 2022). An alternative approach for importance calculation involves impurity measures, such as the Gini index (Louppe et al. 2013). Yet, this method might be unreliable when dealing with varying measurement scales or number of variable categories (Strobl, Boulesteix, Zeileis, et al. 2007). The OOB observations not only contribute to estimating RF error rates but also fulfill the validation criterion in predictive modeling, resembling the principles of  $k$ -fold cross-validation (Hastie, Tibshirani, and Friedman 2009).

The interpretations derived from PI represent **marginal** importances, reflecting the impact of individual variables in predicting outcomes independently (Debeer and Strobl 2020; Strobl, Boulesteix, Kneib, et al. 2008). **Conditional permutation importance** (CPI) extends PI by providing **partial** importances, illustrating the effect of a variable concerning all other variables in the model (Debeer and Strobl 2020; Strobl, Boulesteix, Kneib, et al. 2008). CPI condenses the information from multiple correlated variables into a single variable importance measure, resulting in one variable being assigned higher importance while the importance values of other variables is diminished (Molnar et al. 2022). It is particularly useful in scenarios in datasets with highly correlated variables.

RF's versatility extends to handling mixed variable types, accommodating multi-class problems, and adapting to situations with more variables than observations (Díaz-Uriarte and Alvarez de Andrés 2006). To address imbalanced datasets, the weighted RF approach (C. Chen, Liaw, and Breiman 2004) proposes assigning a heavier penalty for misclassifying minority classes.

### 3.5 Other methods

There are various other supervised predictive modeling methods worth mentioning. Here, two of them are described briefly.

$k$ NN represents a straightforward approach where the model finds the  $k$  observations in the training data that are closest to the observation to be predicted. The value of  $k$  is a user-defined hyperparameter. By assessing the known responses in the training data, the probability of the new observation belonging to a particular class can be directly calculated based on the class distribution in the closest  $k$  neighbor observations. The efficacy of  $k$ NN also hinges on how the distances between observations are computed, known as the distance metric. For instance, the Minkowski distance  $d$  between two observations  $\mathbf{x}_i$  and  $\mathbf{x}_j$  is defined as (Han, Kamber, and Pei 2012; Kuhn and Johnson 2016)

$$d(\mathbf{x}_i, \mathbf{x}_j) = \sqrt[q]{\sum_{p=1}^P |x_{i,p} - x_{j,p}|^q}, \quad (29)$$

where  $q \in \mathbb{R} \geq 1$  and  $P$  represents the number of variables in the data. The

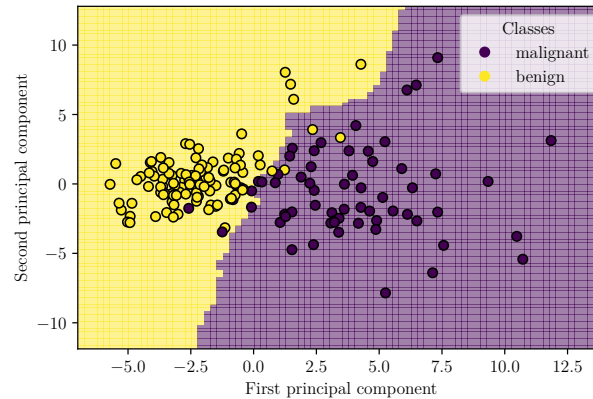


FIGURE 10 An example of the breast cancer data (Wolberg et al. 1995) classification using  $k$ -nearest neighbors ( $k = 10$  with the Euclidean distance metric) (Pedregosa et al. 2011). For clarity, only 30% of the original data were utilized for training. Following a setup similar to Figure 7, the data underwent standardization, and only the first two principal components were considered. The depicted decision boundary showcases the model’s classification of new observations.

Minkowski distance encompasses the common Euclidean (defined as above when  $q = 2$ ) and Manhattan ( $q = 1$ ) distances.

**Neural networks** have served as the foundation for numerous recent applications, such as the multimodal model GPT-4 (OpenAI 2023), capable of processing both natural language via large language models (LLM) and images, and tools designed for generating images based on text, such as DALL-E (Ramesh et al. 2021).

A basic neural network consists of individual **neurons**, organized within a single **hidden layer** or multiple hidden layers, to receive and process input. Each neuron associates weights with its inputs, and these weights are adjusted during the network’s training to generate the appropriate output (Rojas 1996).

Modern neural networks primarily focus on expanding the network’s depth, known as **deep learning**. While this approach demands a substantial amount of training data, it has demonstrated remarkable performance in specific domains, including medical image classifications, translations, text-to-speech applications, and computer vision (Abdou 2022; Aggarwal 2018). Among the prominent deep learning techniques are **convolutional neural networks**, inspired by visual perception, which aim to extract higher-order features from data using convolutional structures (Abdou 2022; Rana and Bhushan 2023). Another approach is **recurrent neural networks**, particularly effective for sequential time series data, which have also shown promise in predicting obesity status (Xue et al. 2018).

## 4 HEALTH DOMAIN APPLICATIONS

In this thesis, the term **prediction** is predominantly used to denote the prediction of a **future** response, based on information procured prior to the time of the response variable. However, it is important to note that this is not a universally accepted definition of prediction in the realm of health-related ML research. ML methodologies have been extensively utilized in tasks that “predict”—or rather, assess—the current situation. This involves evaluating the status of a certain variable based on other variables collected simultaneously. In fact, the examples based on breast cancer data discussed in Chapter 2 and Chapter 3 solely considered the current situation, without any future predictions. At times, the extraction of information that describes the current status can also prove to be beneficial.

Researchers frequently encounter several common pitfalls when utilizing ML. A recent review of sports injury prediction research by Jauhiainen (2023) highlighted issues in assessing the generalization performance of models. One such issue that can lead to biased performance estimates is data leakage, which occurs when information from the testing data inadvertently influences the training phase (Kaufman et al. 2012). Data leakage can emerge when various transformations, variable selection, and/or imputation are applied to the entire dataset at once (Jauhiainen 2023; Kaufman et al. 2012).

Occasionally, neglecting to eliminate the “forbidden” input variables can render the models practically useless. An instance of very problematic data leakage was observed by Kaur, Kumar, and Gupta (2022), who utilized multiple ML methods to determine current weight status. Their study reported impressive results with their gradient boosting classifier, achieving an F-measure of 0.98. The performance of other classifiers was also exceptional. However, these estimates were overly optimistic, as the models had been trained with direct knowledge of how the responses were generated. Specifically, the responses in the data (seven weight classes) were derived directly from the BMI values, calculated as  $BMI = \text{weight}/\text{height}^2$  (Palechor and Hoz Manotas 2019). Despite this, height and weight, along with several other variables, were used as input variables during model training. In this case, the prediction problem can be described as trivial

and could be solved using only height and weight as inputs. The other variables can be seen to have merely added noise to the data.

The health-related domain is vast, and numerous applications have been proposed for a variety of problems. For instance, cancer prediction has caught wide attention. Cancer analysis has been conducted by extracting variables from multiple sources and using these as a structured data basis for the ML models (Sammut et al. 2022) or by leveraging hyperspectral or other imaging techniques, potentially combined with deep learning (Lindholm et al. 2022; Paoli et al. 2022; Prezja et al. 2023). Solutions have also been proposed for detecting different types of skin diseases through user-taken pictures using a regular smartphone camera (Oztel, Yolcu Oztel, and Sahin 2023) and for acute stroke imaging (Sheth et al. 2023). The multi-modality of data—that is, the utilization of different types of data, such as images combined with structured data containing additional information—has also been considered, such as in the context of early diagnosis of Alzheimer’s disease (Diogo, Ferreira, and Prata 2022) and cardiovascular diseases (Amal et al. 2022). Moreover, the relevant data can vary depending on the condition being studied. For instance, ML research related to Parkinson’s disease often utilizes data on a person’s movement and/or voice recordings (Mei, Desrosiers, and Frasnelli 2021).

ML methods have also been employed in predicting diabetes (Alghamdi et al. 2017; De Silva, Jönsson, and Demmer 2019), pressure ulcers (Song et al. 2021; Walther et al. 2022), and asthma and chronic obstructive pulmonary disease (Finkelstein and Jeong 2017; Spathis and Vlamos 2019). Other application areas have included tasks concerning fall risk prediction (Lindberg et al. 2020; Lucero et al. 2019), predicting the outcome of hospitalization in elderly patients (Saarela, Ryyänen, and Äyrämö 2019), sleep disorders (Ha et al. 2023), and mental health, including depression and anxiety, utilizing natural language processing (NLP) models (Le Glaz et al. 2021; Nemesure et al. 2021).

Diverse applications of ML have been proposed for rehabilitation planning and exercise training. For instance, ML has been employed to assist older adults experiencing knee pain (T. Chen and Or 2023); to predict the success of rehabilitation in various hip, knee, or foot injuries (Tschuggnall et al. 2021); and to recommend individually tailored workout activities (Mahyari and Pirolli 2021). Furthermore, sports injury prediction has emerged as another significant area of application (Jauhiainen, Kauppi, Krosshaug, et al. 2022; Jauhiainen, Kauppi, Lepänen, et al. 2021).

This thesis narrows its focus to three specific use cases. The applications included pertain specifically to the areas of obesity and overweight prediction, the development of CRF, and health monitoring. The following sections are dedicated to discussing these application areas in detail.

## 4.1 Obesity and overweight

The prevalence of obesity saw a twofold increase between the years 1980 and 2015 in over 70 countries, and it continues to rise in most other nations (The GBD 2015 Obesity Collaborators 2017). Moreover, the rate of increase in childhood obesity has surpassed that of adult obesity (The GBD 2015 Obesity Collaborators 2017). High BMI has been linked to an increased disease burden, particularly in relation to cardiovascular disease-related deaths (The GBD 2015 Obesity Collaborators 2017). The standard adult BMI cutoff points for overweight and obesity are 25 and 30 kg/m<sup>2</sup>, respectively (The GBD 2015 Obesity Collaborators 2017). Age-specific cutoff points have been established, for instance, for Finnish children and adolescents aged 0–20 years (Saari et al. 2011). However, this seemingly simple metric might not be suitable in some contexts. For instance, while obesity is classified as a disease, these population-based cutoff points do not account for factors such as self-perceived ill-health (Evans and Colls 2009).

In recent years, there has been a significant surge in studies employing a predictive modeling approach to address the issue of obesity and overweight. This section examines some of the recent research not included in Article I.

Zare et al. (2021) utilized LR, decision tree, neural network, and RF models to predict obesity status at the 4th grade, typically around the ages of 9–10 years. The data ( $N = 244,053$ ) used for the prediction task included variables collected in kindergarten at ages 5–6 years. One third of the data were reserved as a held-out test set. The BMI z-score was used alongside race, ethnicity, school meal status, language spoken at home, grade in school, school of attendance, and census block group of residence. The neural network model achieved the highest AUC (0.785), closely followed by LR (0.784) and RF (0.781). The impact of removing the BMI z-score from the input variables was explored, resulting in a dramatic drop in performance, with the AUC of the RF model being only 0.512. When all other variables except the BMI z-score were removed from the models, this had only a minor effect on the performance, with an AUC of 0.782 recorded for the LR model.

Pang et al. (2021) updated their previous study (Pang et al. 2019), included in the review (Article I), to examine other methods in addition to XGBoost, an implementation of gradient boosting. They applied decision tree, Gaussian and Bernoulli naive Bayes methods and LR, neural network, and SVM to predict obesity status at any stage between the ages of 2 and 7 years, based on data collected under 2 years. In total, 102 demographic variables and 54 clinical variables, based on expert clinical assessment, were used. Of the data ( $N = 27,203$ ), 20% were reserved as a held-out test set. The other methods did not enhance the model, and XGBoost was reported as the best model, achieving an AUC of 0.81. When sensitivity was set to 80%, the model's specificity was 63%. The authors concluded that since environmental and genetic factors are known to be associated with obesity, the prediction models might still be improved if further variables of these types are added to the model.

Mondal et al. (2023) utilized RF, LR, neural network, *k*NN, and *k*-means, an unsupervised clustering method, to predict a child's obesity status (normal, overweight, or obese) at the age of 5 years. They explored three distinct research scenarios, each representing different data sources. The first scenario involved data collected from a single well-child visit. All combined 2039 visits of the 224 children available in the data were treated as separate entries during model training. Variables included BMI information, gestational age, birth height, birth weight, and gender. RF emerged as the best-performing method in this scenario. However, as also observed in Article II, predicting a short transition (e.g., from 4 to 5 years), is simpler than predicting a transition from birth to 5 years. This observation was underscored in their study, as accuracy measures for RF were presented separately for different age groups. Reported accuracy ranged between 77% and 100%, with an overall accuracy of 89%. All performance measures were based on using 30% of the data as a separate test set. The second scenario employed data from multiple well-child visits collected from birth up to 2 years. A time-series of BMI values was constructed for each child, utilizing third-degree polynomial interpolation. This method was quite similar to the one seen in Article II, where linear interpolation was also used to fill out the values, making children with various numbers of visits made on different days directly comparable. The overall accuracy for the best model, again RF, was 69%. The third scenario involved using data from multiple random visits under the age of 5 years. The setup in this scenario was otherwise similar to the previous one. The overall accuracy for RF was 89%. As all prediction tasks were implemented as three-class problems, the results are not directly comparable to most other studies.

Regarding the use of deep learning in obesity prediction, a recent review (Ferrerias et al. 2023) did not find any studies that had employed unstructured data in predicting obesity or overweight. However, an updated version of an article mentioned in Article I, Gupta et al. (2022), demonstrated promising results when utilizing structured data with their deep learning long short-term memory (LSTM) network. In a previous study, Xue et al. (2018) leveraged recurrent neural networks for predicting obesity status, using structured activity data. Additionally, many studies have focused on evaluating current obesity status without considering the future. For instance, Rashmi, Umamathy, and Krishnan (2021) used structured data extracted through infrared thermal imaging to determine current obesity status.

## 4.2 Cardiorespiratory fitness development

Low CRF is recognized as being linked with an increased risk of premature mortality from all causes, particularly from cardiovascular disease (American College of Sports Medicine 2018). The gold standard for measuring CRF is the direct measurement of maximal oxygen uptake ( $\text{VO}_2\text{max}$ ), which requires specialized equipment and trained personnel (American College of Sports Medicine 2018;

Liu et al. 2023).

An alternative approach involves estimating  $\text{VO}_2\text{max}$  using non-exercise algorithms. For instance, Tamminen, Laurinen, and Rönning (1999) selected 12 physical and statistical variables that correlated well with the measured oxygen uptake. They utilized regression trees and neural networks in predicting the uptake, and the NN model predictions in particular had high correlations to the measured oxygen uptakes. Recently, Liu et al. (2023) utilized an ML methodology with two distinct gradient-boosting models to predict the current  $\text{VO}_2\text{max}$  value, based on 23 variables in the first model and an additional 26 variables in the second one. The response was defined as continuous, and this method surpassed previously suggested non-exercise methods in performance.

A closely related yet distinct study (Vesterinen et al. 2016) investigated the effectiveness of a submaximal running test (SRT) in predicting running performance and monitoring changes in endurance performance during training. Tracking adaptation to training is important for optimizing training load and facilitating effective recovery. The findings concluded that running speed during the SRT could predict maximal endurance performance, and different stages of the SRT could effectively monitor endurance-training adaptation in recreational endurance runners. The study recognized individual variations in adapting to training loads, highlighting the need for future research to explore the potential productivity of individually tailored training programs.

When considering individual responses to training loads, an individual's genotype significantly influences the extent to which CRF can be improved via training (Bouchard et al. 1999). Similar training loads may yield vastly different outcomes among different individuals, highlighting the inadequacy of the one-size-fits-all approach to training (Bouchard et al. 1999). Therefore, incorporating information about the genotype could be valuable for refining predictive models in this context.

In Article III, the 20MSRT, the most frequently used field test for estimating CRF, was employed. The study aimed to explore the potential of ML in predicting the future development of CRF during adolescence, an application area that remains largely uncharted. In prior research, the primary focus was on investigating treatment procedures and variability in exercise response among adults.

### 4.3 Health monitoring

Health monitoring refers to the tracking and analysis of an individual's health-related data. It can be a continuous process, such as the real-time tracking of a patient's vital signs in a hospital setting, or data can be collected as feasible—for instance, through smart technology—and processed later (Malasinghe, Ramzan, and Dahal 2019). This monitoring can be conducted traditionally with wired sensors or more conveniently with wearable sensors. More sophisticated monitoring solutions can transmit or receive data to remote locations, and even a standard

mobile device, such as a phone, can serve as the processing station or the primary working module (Baig and Gholamhosseini 2013).

A specific example application of monitoring is the automatic detection of falls in elderly individuals. Methodologies for this issue have ranged from analyzing the signal from a wearable accelerometer sensor (Sannino, De Falco, and De Pietro 2015) to a deep learning solution that automatically aims to detect falls from a video feed (Lu et al. 2019).

Applications in health monitoring are predominantly reactive rather than proactive—that is, they aim to detect an event that subsequently requires some form of attention. However, this somewhat technical approach to monitoring some **direct measures of health** may not fully encapsulate the comprehensive concept of health. For instance, according to the World Health Organization (WHO) definition, health is “a state of complete physical, mental and social well-being and not merely the absence of disease or infirmity” (Salomon et al. 2003). Therefore, to portray a more complete picture of health, one should also consider **indirect measures of health**, encompassing self-care, usual activities, interpersonal relations, social functioning, and participation (Salomon et al. 2003).

In Article IV, a new framework for more holistic health monitoring by computing a personal health index was introduced. The methodology in this study was quite distinct from the predictive modeling approaches seen in Article II and Article III. The method was not based on training a model on existing data but was built upon the WHO’s International Classification of Functioning, Disability and Health (ICF) framework (WHO 2001). A primary reason for not employing an ML-based methodology was that capturing reliable data for overall health (i.e., **ground truth**), is practically unattainable. This deficiency makes the utilization of supervised learning methods challenging in this task.

In the realm of predictive modeling and ML, constructing the index holds profound relevance for effective data preparation, addressing the challenges inherent in health data utilization. The index offers dual advantages. First, it provides a concise overview of an individual’s health, serving various purposes, including its incorporation as a response variable in predictive models. Second, the adoption of the ICF framework facilitates structured information collection, enhancing the utility of the data in subsequent predictive modeling tasks.

Over the past two decades, health indices in the literature have included utilizing basic summation approaches to employing techniques such as PCA and multiple correspondence analysis (MCA). Recent advancements have integrated ML and other more sophisticated methodologies into their applications.

Kubik et al. (2002) and Frank et al. (2007) implemented direct summation methods for personal health indices. Kubik et al. (2002) used a Likert-type scale with six questions covering overall health, dietary patterns, and physical activity, scoring from 6 to 24. Frank et al. (2007) queried participants on smoking, drinking, exercise, and diet to construct a similar index. Gallup (2017) employed yes/no questions on general health perceptions, well-restedness, physical discomfort, worry, and sadness, resulting in an index range of  $[0, 100]$  calculated from the mean of valid responses.



Meijer, Kapteyn, and Andreyeva (2011) aimed for an internationally comparable index. They incorporated 25 variables, covering mobility limitations, daily activities, and self-reported health, applying a LISCOMP model integrating factor analysis and regression. Yi et al. (2011) developed cardiovascular, stress, obesity, and management indices using weighted questionnaire and objective measurements. Combining these indices created an overall health index.

Kohn (2012) used MCA to create an index capturing mental health, self-assessment, health issues, disability, accidents, and smoking status. The index range was [1,9], with higher values indicating better health. It allowed flexibility in domain inclusion during the MCA process.

Poterba, Venti, and Wise (2013) utilized PCA to calculate a health index. Their study included 27 health-related queries that encompassed activities of daily living, medical history (e.g., experiences with stroke, diabetes, or cancer), self-reported health, BMI, and other pertinent information. Utilizing the first principal component as their health index, they transformed raw health scores into percentile scores across various ages.

M. K. Kim et al. (2016) introduced a method utilizing a toilet-based system to capture diverse vital signs such as pulse, blood pressure, oxygen saturation, BMI, and ECG readings. Their approach involved constructing five distinct health indices, each focusing on different aspects—namely, heart health, blood parameters, fitness level, muscle condition, and mental health. Although they elaborated on the creation of individual indices, the method of calculating the comprehensive total health index combining these facets was not explicitly detailed.

L. Chen et al. (2016) developed MyPHI, a technique treating health index creation as a predictive modeling problem using soft-label optimization. Employing geriatric medical examination data, this method developed health profiles specifically related to various disease categories. By prioritizing recent health records and employing robust handling of infrequent, incomplete, and sparse data, MyPHI demonstrated superior performance compared to linear SVM and LR models. Disease categories were identified using the WHO's International Classification of Diseases, covering lung, heart, cerebrovascular, diabetes, stomach, colon, liver, pancreas, septicemia, and hypertension.

Lai et al. (2020) introduced an advanced approach for health index construction. Their method fused the Technique for Order Preference by Similarity to Ideal Solution with an independent entropy weighting approach, addressing issues related to missing data in health examination records via tensor decomposition. Acknowledging the complexity of health variables collected during examinations, the study chose nine key health indicators—namely, systolic and diastolic blood pressure, BMI, total/HDL/LDL cholesterol levels, fasting blood glucose, triglyceride levels, and thyrotropin. These indicators were given equal weight in computing the health index, reflecting a holistic yet concise assessment of individual health status across multiple dimensions.

## 4.4 Ethical considerations

Several AI applications have been identified that inadvertently worsen biases against specific populations, often related to gender, ethnicity, and culture. These biases typically stem from skewed training data rather than from deliberate design by developers. Addressing these biases necessitates proactive measures from developers, while the failure to rectify these known issues can intensify biases through feedback loops, further amplifying the use of biased data (I. Y. Chen et al. 2021; Zou and Schiebinger 2018). While the primary focus of this research may not directly engage with these ethical issues, it is important to recognize their relevance in the broader context.

Efforts have been undertaken to establish guidelines in the health-care context to mitigate these issues. For instance, I. Y. Chen et al. (2021) proposed an “ethical pipeline” consisting of five steps to urge developers to consider relevant ethical concerns at each stage of development. They also underscored both visible challenges (e.g., imbalanced and skewed data) and hidden challenges (e.g., group fairness metrics and problem selection bias) in ethical ML model development.

Concrete issues have surfaced in practical settings, as evidenced in medical imaging applications. For example, a model diagnosing diabetic retinopathy demonstrated superior performance in individuals with lighter skin tones compared to those with darker skin tones (Ricci Lara, Echeveste, and Ferrante 2022).

Despite the challenges, the application of ML in the health domain presents a significant opportunity to address existing ethical concerns. For instance, ML can be harnessed to tackle issues encountered by underserved patients, enhance the accessibility of care, institute uniform rules, and assist researchers in rectifying biases prevalent in clinical care (I. Y. Chen et al. 2021).

## 5 SUMMARY OF THE INCLUDED ARTICLES

This thesis delves into the potential of predictive modeling across diverse applications using real-life health data. In this chapter, concise summaries are presented for each of the four articles included.

### 5.1 Article I: Predicting overweight and obesity in later life from childhood data: a review of predictive modeling approaches

Published in *Computational Sciences and Artificial Intelligence in Industry: New Digital Technologies for Solving Future Societal and Economical Challenges*, pp. 203–220, 2021.

**Background:** The global rise in overweight and obesity represents a pressing health concern. The ability to predict the likelihood of future overweight or obesity early in childhood could facilitate timely and effective interventions. Although extensive research has been conducted using explanatory modeling methods, the potential of machine learning for predictive modeling has not been fully explored. Predictive models are validated using unseen examples, providing a more reliable estimate of their real-world performance and generalization capability.

**Objective:** This review aimed to synthesize existing research on childhood overweight or obesity through the lens of predictive modeling.

**Methods:** We conducted two primary literature search cycles in 2018 and 2020, utilizing two databases and Google Scholar. Search terms were carefully selected, and the outcomes of the searches were documented.

**Results and conclusion:** We identified 13 research articles and three review articles relevant to this review. High-performing prediction models typically have a limited prediction timeframe and/or rely on data from later childhood. LR emerged as the most commonly employed technique in developing prediction

models. The variables employed frequently included the child's own anthropometric measurements as well as the mother's weight status and BMI. Contemporary research has also begun to incorporate broader sets of variables and utilize advanced ML techniques. Recurrent neural networks, although scarcely explored, show some promise for predicting overweight and obesity by leveraging their time series prediction capabilities. However, adequate data collection is crucial prior to training such complex models.

## 5.2 Article II: Predicting future overweight and obesity from childhood growth data: a case study

Published in *Computational Sciences and Artificial Intelligence in Industry: New Digital Technologies for Solving Future Societal and Economical Challenges*, pp. 189–201, 2021.

**Background:** The growing prevalence of obesity poses a significant global health challenge, with profound individual and societal impacts due to associated morbidities and mortalities. While factors such as diet and physical activity are recognized contributors to overweight and obesity, predicting individual weight trajectories from early childhood remains complex. The early detection of individuals at risk for developing obesity could significantly enhance preventive measures and interventions in primary health-care settings. Consequently, a robust predictive model could serve as an invaluable tool for health-care professionals.

**Objective:** This study explored the utility of Finnish childhood growth data in constructing ML models to predict future overweight and obesity. To achieve our goal, we utilized the previous review (Article I), which provided appropriate information for reproducing and cross-validating the models on Finnish data.

**Methods and data:** We analyzed body weight and height measurements from 14,197 individuals recorded by health-care specialists in Äänekoski, Finland, from 1986 to 2018. Nine distinct research scenarios were selected based on our literature review. BMI trajectories for each subject were imputed using linear interpolation at 30-day intervals, forming the basis for various predictive models employing LR, SVM, decision trees,  $k$ NN classifiers on principal components, and neural networks.

**Results and conclusion:** The predictive accuracy of our models was on par with, or superior to, those in existing literature. The most effective model, utilizing the SVM approach on Finnish data, achieved an F-measure of 0.73. These findings indicate that the Finnish data may harbor significant patterns conducive to model development. However, incorporating a broader range of early childhood data could potentially refine these predictive models further.

### 5.3 Article III: Precision exercise medicine: predicting unfavourable status and development in the 20-m shuttle run test performance in adolescence with machine learning

Published in *BMJ Open Sport & Exercise Medicine*, 7:e001053, 2021.

**Background:** Precision medicine tailors disease prevention and treatment to individual variability. An analogous concept, precision exercise medicine recognizes the significance of physical activity and CRF in health enhancement. Current research in this field primarily investigates treatment procedures and exercise response variability among adults.

**Objective:** The aim of this study was to evaluate the predictive capabilities of ML, specifically an RF classifier, in predicting adverse future outcomes in the 20MSRT, a prevalent field test for estimating CRF during adolescence.

**Methods and data:** We utilized data from a two-year observational study (2013–2015, mean age  $12.4 \pm 1.3$  years,  $N = 633$  individuals, 50% female) encompassing 48 baseline characteristics derived from questionnaires and objective measurements. These included demographics; physical, psychological, social, and lifestyle factors; anthropometrics; fitness characteristics; physical activity; body composition; and academic scores. The data facilitated predictions for (**Task 1**) unfavorable future 20MSRT status, identifying individuals in the lowest CRF tertile after two years, and (**Task 2**) unfavorable 20MSRT development, pinpointing those with the least progress (lowest tertile) among adolescents below the median baseline 20MSRT.

**Results and conclusion:** The RF classifier demonstrated robust predictive performance for future 20MSRT status (**Task 1**), with an area under the receiver operating characteristic curve (AUC) of 0.83 and 0.76, sensitivity of 80% and 60%, and specificity of 78% and 79% for females and males, respectively. Predictive variables included fitness characteristics, physical activity, academic scores, adiposity, life enjoyment, parental support, social status in school, and perceived fitness. The study's RF classifier successfully identified individuals at risk for unfavorable 20MSRT outcomes, suggesting intervention targets based on a comprehensive profile of 14–20 baseline characteristics. Prediction for future development (**Task 2**) was less accurate, with statistical significance over random levels observed only in females (AUC 0.68 and 0.40 for females and males, respectively). The MATLAB scripts and functions used in this study were made available to advance research in precision exercise medicine.

## 5.4 Article IV: Utilizing the International Classification of Functioning, Disability and Health (ICF) in forming a personal health index

Manuscript, available as a preprint.

**Background:** Personal health indices condense complex health data into a singular value, offering a quick overview of individual health status, serving experts and aiding self-monitoring. This condensed representation holds significance in, for example, rehabilitation for initial health assessment, intervention evaluation, and optimal resource allocation. Challenges in global health data standardization exist due to varying practices. While existing health indices mainly rely on predefined variables, the nuanced nature of personal health necessitates exploring diverse aspects.

**Objective:** We proposed leveraging various data collection methods to aggregate information systematically, enhancing the comprehensiveness of health index calculation. Addressing limitations in existing health index methodologies, we also proposed utilizing the World Health Organization's ICF as the basis for a novel personal health index.

**Methods and data:** To develop and validate this index, data spanning 2013–2019 were collected from a single clinic involving 505 individuals undergoing rehabilitation for various issues. The dataset included questionnaire responses and measurements encompassing the Oswestry low back pain disability questionnaire, EQ-5D-5L generic health questionnaire, mobility and pain level assessments, as well as maximal isometric strength tests. Notably, our approach refrained from prescribing specific sets of variables for inclusion.

**Results and conclusion:** The model underwent two distinct validations, assessing its performance by calculating Pearson correlations between the health index and self-assessed responses for both overall health and maximum pain. The health index outcomes proved to be sensible. In the second validation, Bonferroni correction was applied to enhance the accuracy of the correlation calculations.

Our approach differed significantly from prior methodologies. First, it relied on the ICF framework, accommodating diverse datasets and providing a comprehensive health overview, even with sparse data. Second, it operated independently, deriving the core parameters directly from the ICF hierarchy without relying on training data or predefined parameters. Instead, it computed individualized health indices based on available ICF coded data, focusing on holistic health rather than disease prediction.

Moreover, our model seamlessly handled infrequent health records and missing data, a rarity among existing methodologies. Emphasizing simplicity and reliance on the ICF framework, our approach offered a holistic view of health

across various domains within the ICF. We argue that this health index holds potential for clinical applications, facilitating comparisons between countries and establishing a universal health metric.

## 6 CONCLUSIONS

This thesis encompasses three distinct use cases, each exploring a specific research question. Each case and its corresponding research question are elaborated upon in their respective sections within this chapter. Finally, some limitations and potential future work are discussed.

### 6.1 Potential in obesity and overweight prediction (Q1)

**The first use case** in the thesis diverged into two distinct articles—namely, a review and a case study. To the best of my knowledge, the review (Article I), which mapped existing research on obesity prediction using predictive modeling, was the first and remains the only one to specifically explore this issue. Recently, new reviews (An, J. Shen, and Xiao 2022; Ferreras et al. 2023) have emerged, but their focus deviates from predicting future obesity or overweight status. Instead, they investigate the application of ML methods to obesity and overweight, a considerably broader area. Our review not only explores various methods and their performance in obesity prediction but also aims to understand different research scenarios and their prediction results for the corresponding case study.

The case study (Article II) employed an exclusive dataset provided by the town of Äänekoski encompassing information from over 14,000 individuals. Its primary contribution lies in deploying various predictive modeling techniques to this distinctive Finnish dataset. Notably, it delves into multiple research scenarios, offering deeper insights into the predictability of obesity concerns despite a limited set of variables. This approach sets it apart from prevailing obesity research, often restricted to reporting just one or a few scenarios.

An additional distinctive aspect involves employing interpolation techniques and fully leveraging time series data, enabling direct comparison of data points gathered at different intervals across various individuals. This methodological approach was later also observed in Mondal et al. (2023).

Our case study's findings either matched or surpassed similar studies high-



lighted in the review article. They also align well with previous research indicating that predicting short-term overweight status transitions, such as from ages 2 to 3 (achieving an F-measure of 0.70, sensitivity of 75%, and specificity of 89%), is relatively feasible, although not without challenges. In contrast, predicting longer-term transitions, such as from birth to adolescence (where the F-measure for girls was only 0.41, with sensitivity at 56% and specificity at 60%), proved more challenging.

## 6.2 Potential in cardiorespiratory fitness development prediction (Q2)

The second use case (Article III) aimed to predict the future unfavorable development in CRF among adolescents. The 20MSRT results were used as indicators of CRF. A data-driven ML approach was employed to identify the best predictors of future CRF. The study used an extensive dataset collected from a 2-year longitudinal observational study with 48 variables, including self-assessment questionnaires and noninvasive objective measurements from almost 1,000 children. Novel insights emerged, showcasing the ML's capacity to predict unfavorable future CRF in both girls (AUC 0.83) and boys (AUC 0.76). The variable importance estimates highlighted 14–20 baseline variables linked to future CRF, encompassing factors such as low physical and perceived fitness, high adiposity markers, low physical activity markers, low academic performance, low enjoyment of life, low parental support, and low perceived social status at school.

Despite the challenging nature of predicting unfavorable CRF development in the second task, the achieved AUC (0.68) for girls significantly surpassed the random level. Ten variables emerged with predictive power, echoing a similar set of variables identified in the preceding prediction task.

These results indicate multifaceted influences, where physical, psychological, and social well-being contribute to 20MSRT results in adolescence. This information supplements the existing body of research, which usually focuses on evaluating performance development driven by growth and maturation, resulting in morphological changes. Our findings highlight the added value of assessing an adolescent's holistic well-being over solely relying on their 20MSRT score. This insight advocates a more comprehensive evaluation of individual physical fitness test outcomes, especially in large-scale fitness monitoring systems. Effectively allocating resources for interventions becomes feasible when employing accurate methods tailored for specific individuals.

Furthermore, these findings underscore the potential of ML in identifying candidates for interventions by recognizing data-driven characteristic profiles or phenotypes. Additionally, employing CV techniques helps counter traditional statistical limitations, providing greater insight into the models' generalization capabilities. Robust methodologies are important for adolescents, especially given the scarcity of evidence on criterion references related to physical fitness. This

need is pivotal to avoid over-diagnosis and ensure appropriate measures for the correct individuals.

### 6.3 Developing an ICF-based personal health index and its influence on health assessments and ML (Q3)

The **third use case** (Article IV) focused on addressing the challenges posed by real-world structured health data. The study was rooted in the practical issues encountered by a global network of clinics that collect data in diverse ways. The fundamental approach was to utilize the ICF classification system to circumvent data preparation obstacles that often hinder the practical application of predictive models. The goal was to create a robust personal health index calculation that encapsulates the concept of health in its broadest sense, considering various health aspects. This calculation takes into account several factors, including the ability to manage sparse and infrequent data with missing values and the unification of data from different cultures, regions, and countries. Data quality was also a consideration, meaning that more reliable (i.e., recent or rigorously validated) data carry more weight in the calculation.

The health index was validated using real-world data from individuals undergoing rehabilitation. The results suggest that the proposed model produces valid health index outcomes and is practically applicable. These outcomes can be leveraged in various ways, such as identifying and enhancing factors related to overall health rather than focusing on treating a single symptom, thereby enabling a more holistic view of an individual's health. The proposed model additionally facilitates the creation of comprehensive individual health profiles through a detailed examination of the subcategories within the ICF.

Existing research showcases a variety of methods for computing health indices. However, our model distinguishes itself in the following key aspects:

- It does not depend on predefined variables. Instead, it adapts to any dataset that is compatible with the ICF. However, it is crucial to select relevant ICF codes based on an individual's health aspects, recognizing that different individuals require different ICF codes for a comprehensive health assessment.
- Unlike models that use training data or predefined parameters, it derives its core parameters from the hierarchical structure of the ICF. This allows for the computation of an individual's health index with available ICF-coded data. Each person's health index remains distinct and is not influenced by the health indices of others.
- It concentrates on estimating current health status without predicting disease risks. In line with the WHO's definition of health, it takes into account daily activities, mental health, and social interactions, aligning with the ICF's capacity to capture these aspects.

- Uniquely, it can handle infrequent, longitudinal health records and missing data, unlike most approaches. Its simplicity and reliance on the ICF's hierarchy make it robust against common structured health data issues.

In the context of predictive modeling and ML, the health index offers the advantage of creating a framework to merge datasets from various sources, enhancing data preparation. Ensuring consistent variables across all individuals is vital for robust ML models, and the failure to standardize leads to smaller subsets, undermining predictive models. Moreover, the health index can act as a key optimization target in ML, enabling the use of standardized datasets for predicting individual health. This, for instance, can aid in personalized rehabilitation pathway selection.

In conclusion, our approach provides adaptability to diverse datasets, independence from external factors, a holistic view of health, and robustness against data irregularities. This ensures the generation of outputs for any dataset that meets the minimum ICF-compatible requirements.

## 6.4 Limitations and future work

Cohort studies with a large number of collected variables typically have a relatively low participant count. For various reasons, a significant number of participants may need to be excluded from the final analyses, resulting in a smaller sample size. This reduction impacts the selection of potential predictive modeling methods and validation possibilities. For example, it might be impractical to reserve a separate hold-out dataset for the final validation of the chosen model or to use models that typically require large datasets. Moreover, leveraging the potential of ML methods with more complexity, including deep learning, requires large amounts of data for training, often posing a challenge due to limited data availability. However, augmenting the models with additional variables while leveraging methods with more complexity, such as convolutional neural networks or generative adversarial networks, might enhance their performance (Dogan et al. 2023; Islam and Shamsuddin 2021).

Leveraging previously collected datasets can inspire new ideas for refining data collection. Identifying deficiencies in data collection and post-processing can be a lengthy and slow process. The goal should be to collect data in such a way that it is usable with minimal preparation steps. Ideally, new data should be easily incorporated when refining the model or making predictions with the existing model.

In the field of rehabilitation, an iterative ICF-based rehabilitation cycle, known as Rehab-Cycle, has been formulated for rehabilitation management (Rauch, Cieza, and Stucki 2008). This cycle encompasses the following four key phases: assessment, assignment, intervention, and evaluation. Leveraging the personal health index allows for the creation of a concise numerical measure for both the pre-treatment assessment and the posttreatment evaluation within the Rehab-Cycle.

Moreover, this index can aid in optimizing resource allocation and intervention selection during the assignment and intervention phases. Consequently, it enables a more comprehensive assessment of rehabilitation efficacy and facilitates comparisons among diverse countries and rehabilitation methodologies. With the potential support of ML methods, it could yield insights into best practices for various rehabilitation cases. The future development of the index involves a more rigorous validation in a clinical setting.

Although predictive modeling approaches have gained significant traction in recent years, it is important to acknowledge the vast expanse of uncharted territory that still exists in numerous health domain applications. The potential for discovery and innovation in these areas is immense, and the implications for health-related research can be transformative.

This dissertation has provided an in-depth examination of a select few applications, offering valuable insights and contributing to the growing body of knowledge in this field. Each application holds unique challenges and possibilities, urging further exploration for novel solutions and perspectives.

Although AI and ML tools have made significant progress, their usage remains highly task-specific, demanding substantial manual effort. An **artificial general intelligence** (AGI) solution capable of addressing all problems seems unlikely in the near future and remains a theoretical concept (McLean et al. 2023), despite ongoing efforts in that direction. For example, Y. Shen et al. (2023) proposed using LLMs, such as ChatGPT, as controllers for managing existing AI models. Challenges remain due to the diverse nature of problems, thereby requiring tailored approaches, and because of ethical considerations. Real-world complexities with data and interpretability issues also hinder the development of a comprehensive AI solution. Continuous research and interdisciplinary collaborations with domain experts are essential for addressing these challenges and advancing toward more usable AI and ML solutions.

## YHTEENVETO (SUMMARY IN FINNISH)

Tämä neljän artikkelin väitöskirja tarkastelee olemassa olevien suomalaisten terveysaineistojen potentiaalia yksilöllisen terveyden kehittymiseen liittyvissä kysymyksissä kirjallisuuskatsauksen (Artikkeli I), koneoppimismenetelmien soveltamisen (Artikkeli II ja Artikkeli III) sekä kokonaan uuden laskentamenetelmän luomisen (Artikkeli IV) keinoin.

**Ensimmäinen käyttötapaus** jakautui kahteen erilliseen artikkeliin; kirjallisuuskatsaukseen sekä tapaustutkimukseen. Kirjallisuuskatsauksessa Artikkelissa I tarkasteltiin olemassa olevaa tutkimusta ylipainon ja liikalihavuuden ennustamisessa lasten ja nuorten kohderyhmässä. Katsaukseen päätyneet tutkimukset täyttivät ennustavan mallintamisen kriteerit. Tutkimus tarjosi kattavan kartoituksen alan olemassa olevista tutkimuksista, keskittyen erityisesti tulevan tilan ennustamiseen, toisin kuin ennen tätä katsausta tehdyt sekä viimeaikaiset katsaukset, jotka sisältävät koneoppimisen laajempia sovelluksia ylipainoon ja liikalihavuuteen liittyvissä kysymyksissä. Erilaiset tutkimusasetelmat kirjattiin ylös, jotta niitä voitiin hyödyntää myöhemmin tapaustutkimuksessa suomalaisella aineistolla.

Tapaustutkimus Artikkelissa II hyödynsi Äänekosken kaupungin ainutlaatuista aineistoa, joka sisälsi kasvukäyrätietoja yli 14 000 henkilöltä, ja hyödynsi useita eri ennustavan mallintamisen menetelmiä. Tutkimuksessa hyödynnettiin myös useita tutkimusasetelmia, ja siinä saatiin hyvä yleiskuva ylipainon ja liikalihavuuden ennustettavuudesta suomalaisella aineistolla, vaikka käytössä olikin rajattu määrä muuttujia. Merkittävä piirre oli interpolointitekniikan käyttö, joka mahdollisti eri aikavälein kerättyjen tietopisteiden suoran vertailun yksilöiden välillä. Vastaavaa menetelmää on hyödynnetty hiljattain samassa tarkoituksessa (Mondal et al. 2023).

Tapaustutkimuksen ennustuskyvykkyys eri asetelmille oli linjassa katsauksessa esiinnostettujen tutkimusten kanssa tai jopa ylitti ne. Tutkimus vahvisti myös osaltaan ymmärrettävän ilmiön: lyhyen aikavälin siirtymät ylipainon ennustamisessa, kuten kaksivuotiaasta kolmevuotiaaksi, osoittautui suhteellisen menestykselliseksi (F-mitta 0,70, sensitiivisyys 75 % ja spesifisyys 89 %). Pidemmän aikavälin ennusteet, kuten siirtymä syntymästä nuoruuteen, olivat kuitenkin haastavampia, ja tässä esimerkkitapauksessa tytöillä tulokset olivat vaatimatomat (F-mitta 0,41, sensitiivisyys 56 % ja spesifisyys 60 %).

**Toisen käyttötapauksen** tutkimus Artikkelissa III pyrki ennustamaan kardiorespiratorisen kunnan kehitystä nuorilla aineistolähtöisesti satunnaismetsämenetelmän avulla. Ennustustehtäviä oli kaksi: (**Tehtävä 1**) Heikko kunnan tila; niiden henkilöiden tunnistaminen, jotka ovat alimmassa kuntotertiilissä kahden vuoden seurantajakson lopussa, ja (**Tehtävä 2**) Heikko kunnan kehitys; heikoiten seurantajakson aikana kehittyneiden (alin tertiili tuloksen kehityksessä) henkilöiden tunnistaminen siten, että mukana olivat vain lähtötilanteessa mediaanin alapuolella olevat henkilöt. Kunnan indikaattorina käytettiin 20 metrin viivajuoksu-testin tuloksia. Tutkimus hyödynsi laajaa aineistoa kahden vuoden pituisesta

pitkittäistutkimuksesta, johon osallistui lähes tuhat lasta ja jossa oli 48 muuttujaa. Mallit nostivat esiin ennustusvoimaa sisältäviä kardiorespiratorisen kunnon tulevaisuuden ennustajia, sisältäen sekä itsearviointikyselyitä että noninvasiivisin menetelmin tehtyjä mittauksia. Satunnaismetsä osoitti varsin hyvää ennustuskykyä ennustamalla heikkoa kunnon tilaa (**Tehtävä 1**) sekä tytöillä (AUC 0,83) että pojilla (AUC 0,76), korostaen 14–20 kardiorespiratoriseen kuntoon yhteydessä olevaa muuttujaa, kuten matalaa fyysistä kuntoa, korkeaa rasvaprosenttia sekä sosiaalisia ja koulunkäyntiin liittyviä tekijöitä.

Vaikka heikon kunnon kehityksen ennustaminen (**Tehtävä 2**) oli haastavaa, tyttöjen AUC (0,68) ylitti kuitenkin merkittävästi satunnaisen tason. Mallissa nousi esiin kymmenen ennustusvoimaa sisältävää muuttujaa, muistuttaen aikaisemmassa ennustetehtävässä tunnistettuja muuttujia.

Tulokset viittaavat siihen, että nuorilla viivajuoksutuloksiin vaikuttavat niin fyysinen, psyykinen kuin sosiaalinenkin hyvinvointi. Tämä korostaa kokonaisvaltaisen terveyden arvioinnin merkitystä pelkän viivajuoksutuloksen pisteytyksen sijaan. Tulokset siis puoltavat kattavampaa arviointia, erityisesti suuren mittakaavan kunnonseurantajärjestelmissä. Havainnot tukevat sitä, että koneoppimisella on potentiaalia kunnon kehityksen ennustamisessa ja interventioihin sopivien henkilöiden tunnistamisessa. Lisäksi ristiinvalidointitekniikoilla saadaan parempi ymmärrys mallien yleistymiskyvystä.

**Kolmannen käyttötapauksen** tutkimus Artikkelissa IV keskittyi ratkaisemaan rakenteellisen terveystieteen käsittelyn haasteita. Tutkimuksessa hyödynnettiin ICF-luokitusjärjestelmää, ja tarkoituksena oli huomioida aineiston esikäsittelyn haasteet sekä luoda kattava henkilökohtaisen terveystieteen indeksi. Indeksillä ottaa huomioon erilaiset terveystieteen näkökohdat, kykenee käsittelemään eri kulttuurista peräisin olevaa hajanaista ja monimuotoista aineistoa, sekä antaa suuremman painoarvon tuoreelle ja laadukkaalle tiedolle. Validointi kuntoutuksessa olevien henkilöiden aineistolla osoittaa, että malli tuottaa päteviä terveystieteen arvoja. Toisin kuin useimmat aiemmat menetelmät, malli sopeutuu erilaisiin aineistoihin, kunhan ne ovat yhteensopivia ICF:n kanssa. Mallin ydinparametrit johdetaan suoraan ICF:n hierarkkisesta rakenteesta, eikä se siten tarvitse erillistä koulutusta. Menetelmä keskittyy nykytilanteen terveydentilaan arviointiin, eikä se pyri eri tautien riskien ennustamiseen. Täten se noudattaa WHO:n terveystieteen määritelmää, jossa kokonaisterveyden kannalta olennaisia ovat etenkin päivittäisten toimintojen sujuminen, mielenterveys ja sosiaaliset vuorovaikutukset.

Terveystieteen indeksiä voidaan hyödyntää esimerkiksi kuntoutuksen tehostamisessa, jolloin indeksin avulla on mahdollista luoda numeerisia arvoja arviointeihin ennen hoitoa sekä hoidon jälkeen. Indeksillä voidaan hyödyntää tässä yhteydessä myös resurssien oikeassa kohdistamisessa sekä interventioiden valinnassa. Tämä mahdollistaa kuntoutuksen onnistumisen paremman arvioinnin, eri maiden väliset vertailut sekä tulevaisuudessa mahdollisesti koneoppimismenetelmien avulla parhaiden käytäntöjen tai hoitopolkujen ennustamisen erilaisiin kuntoutustapauksiin.

**Lopuksi**, kohorttitutkimuksissa, joissa on suuri määrä muuttujia, osallistujien määrä on usein melko rajattu ja myös vähenee tutkimuksen edetessä, mikä näkyy käytettävässä aineistossa. Suhteellisen vähäinen aineiston määrä rajoittaa kompleksisempien koneoppimismenetelmien, kuten syväoppimisen, hyödyntämistä. Monipuolisempien muuttujien lisääminen malleihin voi parantaa mallien suorituskykyä. Olemassa olevien tietoaineistojen hyödyntäminen, kuten tässä tutkimuksessa tehtiin, voi myös kehittää aineistonkeruuta ja auttaa huomaamaan niissä olevia puutteita.

Terveystieteiden sovelluksissa on edelleen mahdollisuuksia, joilla on epäilemättä vielä paljon hyödyntämätöntä potentiaalia. Tämä väitöskirja perehtyi monipuolisesti muutamaan valikoituun sovellukseen ja täydentää alan tietämystä. Vaikka tekoäly- ja koneoppimismenetelmät edistyvät koko ajan, vahvan tekoälyn (AGI) saavuttaminen näyttää vielä etäiseltä. Jatkuvasti tehtävä tutkimus ja tieteidenvälinen yhteistyö ovat avainasemassa käytettävien tekoäly- ja koneoppimiseratkaisujen edistämiseksi.

## BIBLIOGRAPHY

- M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng 2015. TensorFlow: large-scale machine learning on heterogeneous systems. [URL: https://www.tensorflow.org/](https://www.tensorflow.org/).
- M. A. Abdou 2022. Literature review: Efficient deep neural networks techniques for medical image analysis. *Neural Computing and Applications* 34 (8), 5791–5812. DOI: 10.1007/s00521-022-06960-9.
- C. C. Aggarwal 2018. *Neural networks and deep learning*. Springer. DOI: 10.1007/978-3-319-94463-0.
- M. A. Ahmad, C. Eckert, and A. Teredesai 2018. Interpretable machine learning in healthcare. In *Proceedings of the 2018 ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*. BCB '18. Association for Computing Machinery, 559–560. DOI: 10.1145/3233547.3233667.
- M. Alghamdi, M. Al-Mallah, S. Keteyian, C. Brawner, J. Ehrman, and S. Sakr 2017. Predicting diabetes mellitus using SMOTE and ensemble machine learning approach: The Henry Ford Exercise Testing (FIT) project. *PLOS ONE* 12 (7), 1–15. DOI: 10.1371/journal.pone.0179805.
- E. Alpaydmn 2014. *Introduction to machine learning*. 3rd ed. MIT Press. ISBN: 978-0-262-02818-9.
- A. Alyass, M. Turcotte, and D. Meyre 2015. From big data analysis to personalized medicine for all: Challenges and opportunities. *BMC Medical Genomics* 8 (1), 33. DOI: 10.1186/s12920-015-0108-y.
- S. Amal, L. Safarnejad, J. A. Omiye, I. Ghazouri, J. H. Cabot, and E. G. Ross 2022. Use of multi-modal data and machine learning to improve cardiovascular disease care. *Frontiers in Cardiovascular Medicine* 9. DOI: 10.3389/fcvm.2022.840262.
- American College of Sports Medicine 2018. *ACSM's guidelines for exercise testing and prescription*. 10th ed. Wolters Kluwer.
- R. An, J. Shen, and Y. Xiao 2022. Applications of artificial intelligence to obesity research: scoping review of methodologies. *Journal of Medical Internet Research* 24 (12), e40589. DOI: 10.2196/40589.



- A. M. Antoniadis, Y. Du, Y. Guendouz, L. Wei, C. Mazo, B. A. Becker, and C. Mooney 2021. Current challenges and future opportunities for XAI in machine learning-based clinical decision support systems: A systematic review. *Applied Sciences* 11 (11). DOI: 10.3390/app11115088.
- S. Äyrämö, T. Kärkkäinen, and K. Majava 2007. Robust refinement of initial prototypes for partitioning-based clustering algorithms. In *Recent Advances in Stochastic Modeling and Data Analysis*. World Scientific, 473–482.
- M. M. Baig and H. Gholamhosseini 2013. Smart health monitoring systems: an overview of design and modeling. *Journal of Medical Systems* 37, 1–14. DOI: 10.1007/s10916-012-9898-z.
- J. Bergstra and Y. Bengio 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research* 13 (10), 281–305. [⟨URL: https://jmlr.org/papers/v13/bergstra12a.html⟩](https://jmlr.org/papers/v13/bergstra12a.html).
- B. Bischl, M. Binder, M. Lang, T. Pielok, J. Richter, S. Coors, J. Thomas, T. Ullmann, M. Becker, A.-L. Boulesteix, D. Deng, and M. Lindauer 2023. Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges. *WIREs Data Mining and Knowledge Discovery* 13 (2), e1484. DOI: 10.1002/widm.1484.
- C. M. Bishop 2006. *Pattern recognition and machine learning*. Springer.
- C. Bouchard, P. An, T. Rice, J. S. Skinner, J. H. Wilmore, J. Gagnon, L. Pérusse, A. S. Leon, and D. C. Rao 1999. Familial aggregation of VO<sub>2</sub>max response to exercise training: results from the HERITAGE Family Study. *Journal of Applied Physiology* 87 (3), 1003–1008. DOI: 10.1152/jappl.1999.87.3.1003.
- R. R. Bouckaert 2003. Choosing between two learning algorithms based on calibrated tests. In *Proceedings of the Twentieth International Conference on International Conference on Machine Learning. ICML'03*. Washington, DC, USA: AAAI Press, 51–58. DOI: 10.5555/3041838.3041845.
- L. Breiman 2001a. Random forests. *Machine Learning* 45 (1), 5–32. DOI: 10.1023/A:1010933404324.
- L. Breiman 2001b. Statistical modeling: The two cultures. *Statistical Science* 16 (3), 199–215. DOI: 10.1214/ss/1009213726.
- J. Brownlee 2020. *Data preparation for machine learning: data cleaning, feature selection, and data transforms in Python*. Machine Learning Mastery.
- A. D. Bull 2011. Convergence rates of efficient global optimization algorithms. *Journal of Machine Learning Research* 12 (88), 2879–2904. [⟨URL: https://jmlr.org/papers/v12/bull11a.html⟩](https://jmlr.org/papers/v12/bull11a.html).
- D. Bzdok, N. Altman, and M. Krzywinski 2018. Points of significance: statistics versus machine learning. *Nature Methods* 15 (4), 233–234. DOI: 10.1038/nmeth.4642.

- A. Callahan and N. H. Shah 2017. Chapter 19 - Machine learning in healthcare. In *Key Advances in Clinical Informatics*. Ed. by A. Sheikh, K. M. Cresswell, A. Wright, and D. W. Bates. Academic Press, 279–291. DOI: 10.1016/B978-0-12-809523-2.00019-4.
- N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer 2002. SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research* 16, 321–357. DOI: 10.1613/jair.953.
- C. Chen, A. Liaw, and L. Breiman 2004. Using random forest to learn imbalanced data. Tech. rep. University of California, Berkeley.
- I. Y. Chen, E. Pierson, S. Rose, S. Joshi, K. Ferryman, and M. Ghassemi 2021. Ethical machine learning in healthcare. *Annual Review of Biomedical Data Science* 4 (1), 123–144. DOI: 10.1146/annurev-biodatasci-092820-114757.
- L. Chen, X. Li, Y. Yang, H. Kurniawati, Q. Z. Sheng, H.-Y. Hu, and N. Huang 2016. Personal health indexing based on medical examinations: A data mining approach. *Decision Support Systems* 81, 54–65. DOI: 10.1016/j.dss.2015.10.008.
- T. Chen and C. K. Or 2023. Perceptions of a machine learning-based lower-limb exercise training system among older adults with knee pain. *Digital Health* 9. DOI: 10.1177/20552076231186069.
- D. Chicco and G. Jurman 2023. The Matthews correlation coefficient (MCC) should replace the ROC AUC as the standard metric for assessing binary classification. *BioData Mining* 16 (4). DOI: 10.1186/s13040-023-00322-4.
- F. Chollet et al. 2015. Keras. [⟨URL: https://keras.io⟩](https://keras.io).
- K. De Silva, D. Jönsson, and R. T. Demmer 2019. A combined strategy of feature selection and machine learning to identify predictors of prediabetes. *Journal of the American Medical Informatics Association* 27 (3), 396–406. DOI: 10.1093/jamia/ocz204.
- D. Debeer and C. Strobl 2020. Conditional permutation importance revisited. *BMC Bioinformatics* 21 (1), 1–30. DOI: 10.1186/s12859-020-03622-2.
- V. Dhar 2013. Data science and prediction. *Communications of the ACM* 56 (12), 64–73. DOI: 10.1145/2500499.
- R. Díaz-Uriarte and S. Alvarez de Andrés 2006. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics* 7 (1), 1–13. DOI: 10.1186/1471-2105-7-3.
- I. D. Dinov 2016. Methodological challenges and analytic opportunities for modeling and interpreting Big Healthcare Data. *GigaScience* 5 (1). DOI: 10.1186/s13742-016-0117-6.
- V. S. Diogo, H. A. Ferreira, and D. Prata 2022. Early diagnosis of Alzheimer’s disease using machine learning: a multi-diagnostic, generalizable approach. *Alzheimer’s Research & Therapy* 14, 107. DOI: 10.1186/s13195-022-01047-y.

- A. Dogan, Y. Li, C. Peter Odo, K. Sonawane, Y. Lin, and C. Liu 2023. A utility-based machine learning-driven personalized lifestyle recommendation for cardiovascular disease prevention. *Journal of Biomedical Informatics* 141, 104342. DOI: 10.1016/j.jbi.2023.104342.
- J. Espregueira-Mendes, R. Barbosa Pereira, and A. Monteiro 2011. Lower limb rehabilitation. In *Orthopedic sports medicine: principles and practice*. Ed. by F. Margheritini and R. Rossi. Milano: Springer, 485–495. DOI: 10.1007/978-88-470-1702-3\_34.
- B. Evans and R. Colls 2009. Measuring fatness, governing bodies: the spatialities of the body mass index (BMI) in anti-obesity politics. *Antipode* 41 (5), 1051–1083. DOI: 10.1111/j.1467-8330.2009.00706.x.
- T. Fawcett 2006. An introduction to ROC analysis. *Pattern Recognition Letters* 27 (8), 861–874. DOI: 10.1016/j.patrec.2005.10.010.
- A. Ferreras, S. Sumalla-Cano, R. Martínez-Licort, I. Elío, K. Tutusaus, T. Prola, J. L. Vidal-Mazón, B. Sahelices, and I. de la Torre Díez 2023. Systematic review of machine learning applied to the prediction of obesity and overweight. *Journal of Medical Systems* 47 (1), 8. DOI: 10.1007/s10916-022-01904-1.
- M. Feurer and F. Hutter 2019. Hyperparameter optimization. In *Automated machine learning: methods, systems, challenges*. Ed. by F. Hutter, L. Kotthoff, and J. Vanschoren. The Springer Series on Challenges in Machine Learning. Springer International Publishing, 3–33.
- D. Filos, A. Triantafyllidis, V. Manolios, K. Livitckaia, J. Claes, R. Buys, V. Cornelissen, E. Kouidi, N. Maglaveras, and I. Chouvarda 2017. Predictive modeling of exercise response in CVD patients under rehabilitation. In *EMBECE & NBC 2017*. Ed. by H. Eskola, O. Väisänen, J. Viik, and J. Hyttinen. Singapore: Springer, 201–204. DOI: 10.1007/978-981-10-5122-7\_51.
- J. Finkelstein and I. c. Jeong 2017. Machine learning approaches to personalize early prediction of asthma exacerbations. *Annals of the New York Academy of Sciences* 1387 (1), 153–165. DOI: 10.1111/nyas.13218.
- E. Frank, J. S. Carrera, L. Elon, and V. S. Hertzberg 2007. Predictors of US medical students' prevention counseling practices. *Preventive Medicine* 44 (1), 76–81. DOI: 10.1016/j.ypmed.2006.07.018.
- J. Futoma, M. Simons, T. Panch, F. Doshi-Velez, and L. A. Celi 2020. The myth of generalisability in clinical research and machine learning in health care. *The Lancet Digital Health* 2 (9), e489–e492. DOI: 10.1016/S2589-7500(20)30186-2.
- Gallup 2017. *Worldwide research methodology and codebook*. Gallup, Inc.
- A. Gelman and J. Hill 2006. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press.
- R. Genuer, J.-M. Poggi, and C. Tuleau-Malot 2010. Variable selection using random forests. *Pattern Recognition Letters* 31 (14), 2225–2236. DOI: 10.1016/j.patrec.2010.03.014.

- M. Ghassemi, T. Naumann, P. Schulam, A. L. Beam, I. Y. Chen, and R. Ranganath 2020. A review of challenges and opportunities in machine learning for health. *AMIA Summits on Translational Science Proceedings 2020*, 191–200.
- J. R. van Ginkel, M. Linting, R. C. A. Rippe, and A. van der Voort 2020. Rebutting existing misconceptions about multiple imputation as a method for handling missing data. *Journal of Personality Assessment* 102 (3), 297–308. DOI: 10.1080/00223891.2018.1530680.
- A. Gnauck 2004. Interpolation and approximation of water quality time series and process identification. *Analytical and Bioanalytical Chemistry* 380, 484–492. DOI: 10.1007/s00216-004-2799-3.
- R. B. Gramacy 2020. *Surrogates: Gaussian process modeling, design and optimization for the applied sciences*. Boca Raton, Florida: Chapman Hall/CRC.
- M. Grandini, E. Bagli, and G. Visani 2020. Metrics for multi-class classification: an overview. DOI: 10.48550/arXiv.2008.05756.
- M. Graziani, L. Dutkiewicz, D. Calvaresi, J. P. Amorim, K. Yordanova, M. Vered, R. Nair, P. H. Abreu, T. Blanke, V. Pulignano, et al. 2022. A global taxonomy of interpretable AI: unifying the terminology for the technical and social sciences. *Artificial Intelligence Review* 56, 3473–3504. DOI: 10.1007/s10462-022-10256-8.
- M. Gupta, T.-L. T. Phan, H. T. Bunnell, and R. Beheshti 2022. Obesity prediction with EHR data: a deep learning approach with interpretable elements. *ACM Transactions on Computing for Healthcare* 3 (3), 1–19. DOI: 10.1145/3506719.
- S. Ha, S. J. Choi, S. Lee, R. H. Wijaya, J. H. Kim, E. Y. Joo, and J. K. Kim 2023. Predicting the risk of sleep disorders using a machine learning-based simple questionnaire: development and validation study. *Journal of Medical Internet Research* 25, e46520. DOI: 10.2196/46520.
- R. Hamon, H. Junklewitz, I. Sanchez, G. Malgieri, and P. De Hert 2022. Bridging the gap between AI and explainability in the GDPR: towards trustworthiness-by-design in automated decision-making. *IEEE Computational Intelligence Magazine* 17 (1), 72–85. DOI: 10.1109/MCI.2021.3129960.
- J. Han, M. Kamber, and J. Pei 2012. *Data mining concepts and techniques*. 3rd ed. Morgan Kaufmann. ISBN: 978-0-12-381479-1.
- T. Hastie, R. Tibshirani, and J. H. Friedman 2009. *The elements of statistical learning: data mining, inference, and prediction*. 2nd ed. Springer. DOI: 10.1007/978-0-387-21606-5.
- J. Huang and C. Ling 2005. Using AUC and accuracy in evaluating learning algorithms. *IEEE Transactions on Knowledge and Data Engineering* 17 (3), 299–310. DOI: 10.1109/TKDE.2005.50.
- T. Hujanen, M. Peltola, U. Häkkinen, and M. Pekurinen 2008. Miesten ja naisten terveystutkimus ikäryhmittäin 2006. *Stakes / Työpapereita* 37. [⟨URL: https://urn.fi/URN:NBN:fi-fe201204193907⟩](https://urn.fi/URN:NBN:fi-fe201204193907).

- R. Iniesta, D. Stahl, and P. McGuffin 2016. Machine learning, statistical learning and the future of biological research in psychiatry. *Psychological Medicine* 46 (12), 2455–2465. DOI: 10.1017/S0033291716001367.
- M. Islam and R. Shamsuddin 2021. Machine learning to promote health management through lifestyle changes for hypertension patients. *Array* 12, 100090. DOI: 10.1016/j.array.2021.100090.
- S. Jauhiainen 2023. Potential of predictive modeling methods for individual response: applications and guidelines for sports sciences. PhD thesis. University of Jyväskylä. [URL: https://urn.fi/URN:ISBN:978-951-39-9697-0](https://urn.fi/URN:ISBN:978-951-39-9697-0).
- S. Jauhiainen, J.-P. Kauppi, T. Krosshaug, R. Bahr, J. Bartsch, and S. Äyrämö 2022. Predicting ACL injury using machine learning on data from an extensive screening test battery of 880 female elite athletes. *The American Journal of Sports Medicine* 50 (11), 2917–2924. DOI: 10.1177/03635465221112095.
- S. Jauhiainen, J.-P. Kauppi, M. Leppänen, K. Pasanen, J. Parkkari, T. Vasankari, P. Kannus, and S. Äyrämö 2021. New machine learning approach for detection of injury risk factors in young team sport athletes. *International Journal of Sports Medicine* 42 (2), 175–182. DOI: 10.1055/a-1231-5304.
- F. Jiang, Y. Jiang, H. Zhi, Y. Dong, H. Li, S. Ma, Y. Wang, Q. Dong, H. Shen, and Y. Wang 2017. Artificial intelligence in healthcare: past, present and future. *Stroke and Vascular Neurology* 2 (4). DOI: 10.1136/svn-2017-000101.
- D. R. Jones, M. Schonlau, and W. J. Welch 1998. Efficient global optimization of expensive black-box functions. *Journal of Global Optimization* 13, 455–492. DOI: 10.1023/A:1008306431147.
- M. Kantardzic 2020. *Data mining: concepts, models, methods, and algorithms*. 3rd ed. IEEE Press & Wiley.
- S. Kaufman, S. Rosset, C. Perlich, and O. Stitelman 2012. Leakage in data mining: formulation, detection, and avoidance. *ACM Transactions on Knowledge Discovery from Data* 6 (4), 15. DOI: 10.1145/2382577.2382579.
- R. Kaur, R. Kumar, and M. Gupta 2022. Predicting risk of obesity and meal planning to reduce the obese in adulthood using artificial intelligence. *Endocrine* 78 (3), 458–469. DOI: 10.1007/s12020-022-03215-4.
- M. J. Khoury and J. P. A. Ioannidis 2014. Big data meets public health. *Science* 346 (6213), 1054–1055. DOI: 10.1126/science.aaa2709.
- J.-H. Kim 2009. Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis* 53 (11), 3735–3745. DOI: 10.1016/j.csda.2009.04.009.
- M. K. Kim, H. T. Jung, S. D. Kim, and H. J. La 2016. A personal health index system with IoT devices. In *2016 IEEE International Conference on Mobile Services (MS)*, 174–177. DOI: 10.1109/MobServ.2016.34.
- C. Kingsford and S. L. Salzberg 2008. What are decision trees? *Nature Biotechnology* 26 (9), 1011–1013. DOI: 10.1038/nbt0908-1011.

- N. Kleinman, S. Abouzaid, L. Andersen, Z. Wang, and A. Powers 2014. Cohort analysis assessing medical and nonmedical cost associated with obesity in the workplace. *Journal of Occupational and Environmental Medicine* 56 (2), 161–170. DOI: 10.1097/JOM.0000000000000099.
- K. Koh, S.-J. Kim, and S. Boyd 2007. An interior-point method for large-scale l1-regularized logistic regression. *Journal of Machine Learning Research* 8, 1519–1555. [⟨URL: https://jmlr.org/papers/v8/koh07a.html⟩](https://jmlr.org/papers/v8/koh07a.html).
- R. Kohavi 1995. A study of cross-validation and bootstrap for accuracy estimation and model selection. In *Proceedings of the Fourteenth International Joint Conference on Artificial Intelligence*. Vol. 2. International Joint Conferences on Artificial Intelligence, 1137–1145. [⟨URL: https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf⟩](https://www.ijcai.org/Proceedings/95-2/Papers/016.pdf).
- J. L. Kohn 2012. What is health? A multiple correspondence health index. *Eastern Economic Journal* 38 (2), 223–250. DOI: 10.1057/ej.2011.5.
- D. Krstajic, L. J. Buturovic, D. E. Leahy, and S. Thomas 2014. Cross-validation pitfalls when selecting and assessing regression and classification models. *Journal of Cheminformatics* 6 (1), 10. DOI: 10.1186/1758-2946-6-10.
- H. M. Krumholz 2014. Big data and new knowledge in medicine: the thinking, training, and tools needed for a learning health system. *Health Affairs* 33 (7), 1163–1170. DOI: 10.1377/hlthaff.2014.0053.
- M. Y. Kubik, L. A. Lytle, P. J. Hannan, M. Story, and C. L. Perry 2002. Food-related beliefs, eating behavior, and classroom food practices of middle school teachers. *Journal of School Health* 72 (8), 339–345. DOI: 10.1111/j.1746-1561.2002.tb07921.x.
- M. Kuhn and K. Johnson 2016. *Applied predictive modeling*. Springer. DOI: 10.1007/978-1-4614-6849-3.
- A. L’Heureux, K. Grolinger, H. F. Elyamany, and M. A. M. Capretz 2017. Machine learning with big data: challenges and approaches. *IEEE Access* 5, 7776–7797. DOI: 10.1109/ACCESS.2017.2696365.
- G. Lai, D. Yu, S. Zhang, Z. Wei, and X. Sun 2020. Personal health index based on residential health examination. In *Advanced Data Mining and Applications*. Ed. by X. Yang, C.-D. Wang, M. S. Islam, and Z. Zhang. Springer International Publishing, 569–581. DOI: 10.1007/978-3-030-65390-3\_43.
- A. Le Glaz, Y. Haralambous, D.-H. Kim-Dufor, P. Lenca, R. Billot, T. C. Ryan, J. Marsh, J. DeVyllder, M. Walter, S. Berrouiguet, and C. Lemey 2021. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research* 23 (5), e15708. DOI: 10.2196/15708.

- D. S. Lindberg, M. Prosperi, R. I. Bjarnadottir, J. Thomas, M. Crane, Z. Chen, K. Shear, L. M. Solberg, U. A. Snigurska, Y. Wu, Y. Xia, and R. J. Lucero 2020. Identification of important factors in an inpatient fall risk prediction model to improve the quality of care using EHR and electronic administrative data: A machine-learning approach. *International Journal of Medical Informatics* 143, 104272. DOI: 10.1016/j.ijmedinf.2020.104272.
- V. Lindholm, A.-M. Raita-Hakola, L. Annala, M. Salmivuori, L. Jeskanen, H. Saari, S. Koskenmies, S. Pitkänen, I. Pölönen, K. Isoherranen, and A. Ranki 2022. Differentiating malignant from benign pigmented or non-pigmented skin tumours – a pilot study on 3D hyperspectral imaging of complex skin surfaces and convolutional neural networks. *Journal of Clinical Medicine* 11 (7). DOI: 10.3390/jcm11071914.
- Y. Liu, J. Herrin, C. Huang, R. Khera, L. S. Dhingra, W. Dong, B. J. Mortazavi, H. M. Krumholz, and Y. Lu 2023. Nonexercise machine learning models for maximal oxygen uptake prediction in national population surveys. *Journal of the American Medical Informatics Association* 30 (5), 943–952. DOI: 10.1093/jamia/ocad035.
- P. R. Lorenzo, J. Nalepa, M. Kawulok, L. S. Ramos, and J. R. Pastor 2017. Particle swarm optimization for hyper-parameter selection in deep neural networks. In *Proceedings of the Genetic and Evolutionary Computation Conference. GECCO '17*. Association for Computing Machinery, 481–488. DOI: 10.1145/3071178.3071208.
- G. Louppe, L. Wehenkel, A. Suter, and P. Geurts 2013. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems*. Ed. by C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger. Vol. 26. Curran Associates, Inc., 431–439.
- N. Lu, Y. Wu, L. Feng, and J. Song 2019. Deep learning for fall detection: three-dimensional CNN combined with LSTM on video kinematic data. *IEEE Journal of Biomedical and Health Informatics* 23 (1), 314–323. DOI: 10.1109/JBHI.2018.2808281.
- R. J. Lucero, D. S. Lindberg, E. A. Fehlberg, R. I. Bjarnadottir, Y. Li, J. P. Cimiotti, M. Crane, and M. Prosperi 2019. A data-driven and practice-based approach to identify risk factors associated with hospital-acquired falls: Applying manual and semi- and fully-automated methods. *International Journal of Medical Informatics* 122, 63–69. DOI: 10.1016/j.ijmedinf.2018.11.006.
- J. Ludwig and S. Mullainathan 2023. Machine learning as a tool for hypothesis generation. Working Paper 31017. National Bureau of Economic Research. DOI: 10.3386/w31017.
- S. M. Lundberg and S.-I. Lee 2017. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*. [URL: https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html](https://proceedings.neurips.cc/paper/2017/hash/8a20a8621978632d76c43dfd28b67767-Abstract.html).

- S. J. MacEachern and N. D. Forkert 2021. Machine learning for precision medicine. *Genome* 64 (4), 416–425. DOI: 10.1139/gen-2020-0131.
- A. Mahyari and P. Pirolli 2021. Physical exercise recommendation and success prediction using interconnected recurrent neural networks. In 2021 IEEE International Conference on Digital Health (ICDH), 148–153. DOI: 10.1109/ICDH52753.2021.00027.
- L. P. Malasinghe, N. Ramzan, and K. Dahal 2019. Remote patient monitoring: a comprehensive study. *Journal of Ambient Intelligence and Humanized Computing* 10, 57–76. DOI: 10.1007/s12652-017-0598-x.
- MathWorks 2023. Bayesian optimization algorithm. The MathWorks, Inc. Accessed: 31.1.2023. [URL: https://se.mathworks.com/help/stats/bayesian-optimization-algorithm.html](https://se.mathworks.com/help/stats/bayesian-optimization-algorithm.html).
- S. McLean, G. J. M. Read, J. Thompson, C. Baber, N. A. Stanton, and P. M. Salmon 2023. The risks associated with artificial general intelligence: a systematic review. *Journal of Experimental & Theoretical Artificial Intelligence* 35 (5), 649–663. DOI: 10.1080/0952813X.2021.1964003.
- J. Mei, C. Desrosiers, and J. Frasnelli 2021. Machine learning for the diagnosis of Parkinson’s disease: a review of literature. *Frontiers in Aging Neuroscience* 13. DOI: 10.3389/fnagi.2021.633752.
- E. Meijer, A. Kapteyn, and T. Andreyeva 2011. Internationally comparable health indices. *Health Economics* 20 (5), 600–619. DOI: 10.1002/hec.1620.
- T. Miller 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence* 267, 1–38. DOI: 10.1016/j.artint.2018.07.007.
- T. Miller 2023. Explainable AI is dead, long live explainable AI! Hypothesis-driven decision support using evaluative AI. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency. FAccT ’23*. Association for Computing Machinery, 333–342. DOI: 10.1145/3593013.3594001.
- J. Milton and J. C. Arnold 1990. *Introduction to probability and statistics: principles and applications for engineering and the computing sciences*. 2nd ed. McGraw–Hill Publishing Company. ISBN: 0-07-100812-8.
- E. Mirkes, T. Coats, J. Levesley, and A. Gorban 2016. Handling missing data in large healthcare dataset: A case study of unknown trauma outcomes. *Computers in Biology and Medicine* 75, 203–216. DOI: 10.1016/j.combiomed.2016.06.004.
- C. Molnar, G. König, J. Herbinger, T. Freiesleben, S. Dandl, C. A. Scholbeck, G. Casalicchio, M. Grosse-Wentrup, and B. Bischl 2022. General pitfalls of model-agnostic interpretation methods for machine learning models. In *xxAI - Beyond Explainable AI: International Workshop, Held in Conjunction with ICML 2020, July 18, 2020, Vienna, Austria, Revised and Extended Papers*. Ed. by A. Holzinger, R. Goebel, R. Fong, T. Moon, K.-R. Müller, and W. Samek. Springer International Publishing, 39–68. DOI: 10.1007/978-3-031-04083-2\_4.



- P. K. Mondal, K. H. Foysal, B. A. Norman, and L. S. Gittner 2023. Predicting childhood obesity based on single and multiple well-child visit data using machine learning classifiers. *Sensors* 23 (2), 759. DOI: 10.3390/s23020759.
- S. J. Mooney and V. Pejaver 2018. Big data in public health: Terminology, machine learning, and privacy. *Annual Review of Public Health* 39, 95–112. DOI: 10.1146/annurev-publhealth-040617-014208.
- D. S. Moore, G. P. McCabe, and B. A. Craig 2009. Introduction to the practice of statistics. 6th ed. W.H. Freeman and Company. ISBN: 978-1-4292-1623-4.
- M. D. Nemesure, M. V. Heinz, R. Huang, and N. C. Jacobson 2021. Predictive modeling of depression and anxiety using electronic health records and a novel machine learning approach with artificial intelligence. *Scientific reports* 11, 1980. DOI: 10.1038/s41598-021-81368-4.
- W. S. Noble 2006. What is a support vector machine? *Nature Biotechnology* 24 (12), 1565–1567. DOI: 10.1038/nbt1206-1565.
- P. J. Olver 2006. On multivariate interpolation. *Studies in Applied Mathematics* 116 (2), 201–240. DOI: 10.1111/j.1467-9590.2006.00335.x.
- OpenAI 2023. GPT-4 technical report. arXiv. DOI: 10.48550/arXiv.2303.08774.
- M. Oquendo, E. Baca-Garcia, A. Artés-Rodríguez, F. Perez-Cruz, H. Galfalvy, H. Blasco-Fontecilla, D. Madigan, and N. Duan 2012. Machine learning and data mining: strategies for hypothesis generation. *Molecular Psychiatry* 17 (10), 956–959. DOI: 10.1038/mp.2011.173.
- I. Oztel, G. Yolcu Oztel, and V. H. Sahin 2023. Deep learning-based skin diseases classification using smartphones. *Advanced Intelligent Systems* 5 (12), 2300211. DOI: 10.1002/aisy.202300211.
- F. M. Palechor and A. de la Hoz Manotas 2019. Dataset for estimation of obesity levels based on eating habits and physical condition in individuals from Colombia, Peru and Mexico. *Data in Brief* 25, 104344. DOI: 10.1016/j.dib.2019.104344.
- X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino 2019. Understanding early childhood obesity via interpretation of machine learning model predictions. In 2019 18th IEEE International Conference On Machine Learning And Applications (ICMLA), 1438–1443. DOI: 10.1109/ICMLA.2019.00235.
- X. Pang, C. B. Forrest, F. Lê-Scherban, and A. J. Masino 2021. Prediction of early childhood obesity with machine learning and electronic health record data. *International Journal of Medical Informatics* 150, 104454. DOI: 10.1016/j.ijmedinf.2021.104454.
- J. Paoli, I. Pölönen, M. Salmivuori, J. Räsänen, O. Zaar, S. Polesie, S. Koskenmies, S. Pitkänen, M. Övermark, K. Isoherranen, S. Juteau, A. Ranki, M. Grönroos, and N. Neittaanmäki 2022. Hyperspectral imaging for non-invasive diagnostics of melanocytic lesions. *Acta dermato-venereologica* 102. DOI: 10.2340/actadv.v102.2045.

- A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimshe, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala 2019. PyTorch: an imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 8024–8035.
- S. N. Payrovnaziri, Z. Chen, P. Rengifo-Moreno, T. Miller, J. Bian, J. H. Chen, X. Liu, and Z. He 2020. Explainable artificial intelligence models using real-world electronic health record data: A systematic scoping review. *Journal of the American Medical Informatics Association* 27 (7), 1173–1185. DOI: 10.1093/jamia/ocaa053.
- F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay 2011. Scikit-learn: machine learning in Python. *Journal of Machine Learning Research* 12, 2825–2830. [URL: https://jmlr.org/papers/v12/pedregosa11a.html](https://jmlr.org/papers/v12/pedregosa11a.html).
- J. M. Poterba, S. F. Venti, and D. A. Wise 2013. Health, education, and the post-retirement evolution of household assets. Working Paper 18695. National Bureau of Economic Research. DOI: 10.3386/w18695.
- F. Prezja, S. Äyrämö, I. Pölönen, T. Ojala, S. Lahtinen, P. Ruusuvoori, and T. Kuopio 2023. Improved accuracy in colorectal cancer tissue decomposition through refinement of established deep learning solutions. *Scientific Reports* 13, 15879. DOI: 10.1038/s41598-023-42357-x.
- P. Probst and A.-L. Boulesteix 2018. To tune or not to tune the number of trees in random forest. *The Journal of Machine Learning Research* 18 (181), 1–18. [URL: https://jmlr.org/papers/v18/17-269.html](https://jmlr.org/papers/v18/17-269.html).
- P. Probst, A.-L. Boulesteix, and B. Bischl 2019. Tunability: importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research* 20 (53), 1–32. [URL: https://jmlr.org/papers/v20/18-444.html](https://jmlr.org/papers/v20/18-444.html).
- P. Probst, M. N. Wright, and A.-L. Boulesteix 2019. Hyperparameters and tuning strategies for random forest. *WIREs Data Mining and Knowledge Discovery* 9 (3), e1301. DOI: 10.1002/widm.1301.
- J. R. Quinlan 1986. Induction of decision trees. *Machine Learning* 1, 81–106. DOI: 10.1007/BF00116251.
- A. Ramesh, M. Pavlov, G. Goh, S. Gray, C. Voss, A. Radford, M. Chen, and I. Sutskever 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*. Ed. by M. Meila and T. Zhang. Vol. 139. *Proceedings of Machine Learning Research*. PMLR, July 2021, 8821–8831. [URL: https://proceedings.mlr.press/v139/ramesh21a.html](https://proceedings.mlr.press/v139/ramesh21a.html).

- M. Rana and M. Bhushan 2023. Machine learning and deep learning approach for medical image analysis: diagnosis to detection. *Multimedia Tools and Applications* 82, 26731–26769. DOI: 10.1007/s11042-022-14305-w.
- R. Rashmi, S. Umapathy, and P. T. Krishnan 2021. Thermal imaging method to evaluate childhood obesity based on machine learning techniques. *International Journal of Imaging Systems and Technology* 31 (3), 1752–1768. DOI: 10.1002/ima.22572.
- C. E. Rasmussen and C. K. Williams 2006. *Gaussian processes for machine learning*. MIT Press. ISBN: 0-262-18253-X.
- A. Rauch, A. Cieza, and G. Stucki 2008. How to apply the International Classification of Functioning, Disability and Health (ICF) for rehabilitation management in clinical practice. *European Journal of Physical and Rehabilitation Medicine* 44 (3), 329–342. ISSN: 1973-9087.
- K. Reini and J. Honkatukia 2016. Hyvä hoito kannattaa: Diabeteksen ennaltaehkäisy ja tehostetun hoidon kansantaloudellinen vaikuttavuus. Vol. 206. Vaasan yliopiston julkaisuja, selvityksiä ja raportteja. University of Vaasa. (URL: <https://urn.fi/URN:ISBN:978-952-476-673-9>).
- M. T. Ribeiro, S. Singh, and C. Guestrin 2016. “Why should I trust you?”: Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. KDD ’16. Association for Computing Machinery, 1135–1144. DOI: 10.1145/2939672.2939778.
- M. A. Ricci Lara, R. Echeveste, and E. Ferrante 2022. Addressing fairness in artificial intelligence for medical imaging. *Nature Communications* 13 (1), 4581. DOI: 10.1038/s41467-022-32186-3.
- D. M. Roden 2016. Cardiovascular pharmacogenomics: current status and future directions. *Journal of Human Genetics* 61, 79–85. DOI: 10.1038/jhg.2015.78.
- P. Rodríguez, M. A. Bautista, J. González, and S. Escalera 2018. Beyond one-hot encoding: Lower dimensional target embedding. *Image and Vision Computing* 75, 21–31. DOI: 10.1016/j.imavis.2018.04.004.
- F. Rodríguez-Torres, J. A. Carrasco-Ochoa, and J. F. Martínez-Trinidad 2021. Experimental comparison of oversampling methods for mixed datasets. In *Pattern Recognition*. Ed. by E. Roman-Rangel, Á. F. Kuri-Morales, J. F. Martínez-Trinidad, J. A. Carrasco-Ochoa, and J. A. Olvera-López. Springer International Publishing, 78–88. DOI: 10.1007/978-3-030-77004-4\_8.
- R. Rojas 1996. *Neural networks: a systematic introduction*. Springer.
- R. Ross, B. H. Goodpaster, L. G. Koch, M. A. Sarzynski, W. M. Kohrt, N. M. Johannsen, J. S. Skinner, A. Castro, B. A. Irving, R. C. Noland, L. M. Sparks, G. Spielmann, A. G. Day, W. Pitsch, W. G. Hopkins, and C. Bouchard 2019. Precision exercise medicine: understanding exercise response variability. *British Journal of Sports Medicine* 53 (18), 1141–1153. DOI: 10.1136/bjsports-2018-100328.

- C. Rudin 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence* 1 (5), 206–215. DOI: 10.1038/s42256-019-0048-x.
- C. Rudin, C. Chen, Z. Chen, H. Huang, L. Semenova, and C. Zhong 2022. Interpretable machine learning: Fundamental principles and 10 grand challenges. *Statistics Surveys* 16, 1–85. DOI: 10.1214/21-SS133.
- T. Ruohonen, R. Kuoremäki, K. Kaasalainen, and O. Kilpi 2018. Asiakas on-line -hanke: loppuraportti. Tech. rep. University of Jyväskylä. [URL: https://urn.fi/URN:ISBN:978-951-39-7369-8](https://urn.fi/URN:ISBN:978-951-39-7369-8).
- M. Saarela, O.-P. Ryyänen, and S. Äyrämö 2019. Predicting hospital associated disability from imbalanced data using supervised learning. *Artificial Intelligence in Medicine* 95, 88–95. DOI: 10.1016/j.artmed.2018.09.004.
- A. Saari, U. Sankilampi, M.-L. Hannila, V. Kiviniemi, K. Kesseli, and L. Dunkel 2011. New Finnish growth references for children and adolescents aged 0 to 20 years: length/height-for-age, weight-for-length/height, and body mass index-for-age. *Annals of Medicine* 43 (3), 235–248. DOI: 10.3109/07853890.2010.515603.
- K. L. Sainani 2014. Explanatory versus predictive modeling. *PM&R* 6 (9), 841–844. DOI: 10.1016/j.pmrj.2014.08.941.
- J. A. Salomon, C. D. Mathers, S. Chatterji, R. Sadana, T. B. Üstün, and C. J. Murray 2003. Quantifying individual levels of health: definitions, concepts and measurement issues. In *Health Systems Performance Assessment: Debate, Methods, and Empiricism*. Geneva: World Health Organization, 301–318. ISBN: 92-4-156245-5.
- S.-J. Sammut, M. Crispin-Ortiz, S.-F. Chin, E. Provenzano, H. A. Bardwell, W. Ma, W. Cope, A. Dariush, S.-J. Dawson, J. E. Abraham, et al. 2022. Multi-omic machine learning predictor of breast cancer therapy response. *Nature* 601 (7894), 623–629. DOI: 10.1038/s41586-021-04278-5.
- G. Sannino, I. De Falco, and G. De Pietro 2015. A supervised approach to automatically extract a set of rules to support fall detection in an mHealth system. *Applied Soft Computing* 34, 205–216. DOI: 10.1016/j.asoc.2015.04.060.
- F. Schiavone and M. Ferretti 2021. The FutureS of healthcare. *Futures* 134, 102849. DOI: 10.1016/j.futures.2021.102849.
- B. Shahriari, K. Swersky, Z. Wang, R. P. Adams, and N. de Freitas 2016. Taking the human out of the loop: a review of Bayesian optimization. *Proceedings of the IEEE* 104 (1), 148–175. DOI: 10.1109/JPROC.2015.2494218.
- Y. Shen, K. Song, X. Tan, D. Li, W. Lu, and Y. Zhuang 2023. HuggingGPT: solving AI tasks with ChatGPT and its friends in Hugging Face. DOI: 10.48550/arXiv.2303.17580.
- S. A. Sheth, L. Giancardo, M. Colasurdo, V. M. Srinivasan, A. Niktabe, and P. Kan 2023. Machine learning and acute stroke imaging. *Journal of NeuroInterventional Surgery* 15 (2), 195–199. DOI: 10.1136/neurintsurg-2021-018142.

- G. Shmueli 2010. To explain or to predict? *Statistical Science* 25 (3), 289–310. DOI: 10.1214/10-STS330.
- N. Shrestha, Z. Pedisic, S. Neil-Sztramko, K. T. Kukkonen-Harjula, and V. Hermans 2016. The impact of obesity in the workplace: a review of contributing factors, consequences and potential solutions. *Current Obesity Reports* 5, 344–360. DOI: 10.1007/s13679-016-0227-6.
- E. Slade and M. G. Naylor 2020. A fair comparison of tree-based and parametric methods in multiple imputation by chained equations. *Statistics in Medicine* 39 (8), 1156–1166. DOI: 10.1002/sim.8468.
- J. Snoek, H. Larochelle, and R. P. Adams 2012. Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems 25 (NIPS 2012)*. Ed. by F. Pereira, C. Burges, L. Bottou, and K. Weinberger, 2951–2959. [URL: https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html](https://proceedings.neurips.cc/paper/2012/hash/05311655a15b75fab86956663e1819cd-Abstract.html).
- W. Song, M.-J. Kang, L. Zhang, W. Jung, J. Song, D. W. Bates, and P. C. Dykes 2021. Predicting pressure injury using nursing assessment phenotypes and machine learning methods. *Journal of the American Medical Informatics Association* 28 (4), 759–765. DOI: 10.1093/jamia/ocaa336.
- D. Spathis and P. Vlamos 2019. Diagnosing asthma and chronic obstructive pulmonary disease with machine learning. *Health Informatics Journal* 25 (3), 811–827. DOI: 10.1177/1460458217723169.
- C. Strobl, A.-L. Boulesteix, T. Kneib, T. Augustin, and A. Zeileis 2008. Conditional variable importance for random forests. *BMC Bioinformatics* 9, 1–11. DOI: 10.1186/1471-2105-9-307.
- C. Strobl, A.-L. Boulesteix, A. Zeileis, and T. Hothorn 2007. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics* 8, 25. DOI: 10.1186/1471-2105-8-25.
- S. Tamminen, P. Laurinen, and J. Röning 1999. Comparing regression trees with neural networks in aerobic fitness approximation. In *Proceedings of the international computing sciences conference symposium on advances in intelligent data analysis*, Rochester, NY, June 22–25, 414–419.
- P.-N. Tan, M. Steinbach, V. Kumar, and A. Karpatne 2013. *Introduction to data mining*. 1st ed. Pearson. ISBN: 978-1-292-02615-2.
- F. Tang and H. Ishwaran 2017. Random forest missing data algorithms. *Statistical Analysis and Data Mining: The ASA Data Science Journal* 10 (6), 363–377. DOI: 10.1002/sam.11348.
- The GBD 2015 Obesity Collaborators 2017. Health effects of overweight and obesity in 195 countries over 25 Years. *New England Journal of Medicine* 377, 13–27. DOI: 10.1056/NEJMoa1614362.

- G. R. Tomkinson, J. J. Lang, M. S. Tremblay, M. Dale, A. G. LeBlanc, K. Belanger, F. B. Ortega, and L. Léger 2017. International normative 20 m shuttle run values from 1 142 026 children and youth representing 50 countries. *British Journal of Sports Medicine* 51 (21), 1545–1554. DOI: 10.1136/bjsports-2016-095987.
- M. Tschuggnall, V. Grote, M. Pirchl, B. Holzner, G. Rumpold, and M. J. Fischer 2021. Machine learning approaches to predict rehabilitation success based on clinical and patient-reported outcome measures. *Informatics in Medicine Unlocked* 24. DOI: 10.1016/j.imu.2021.100598.
- C. Tuena, M. Semonella, J. Fernández-Álvarez, D. Colombo, and P. Cipresso 2020. Predictive precision medicine: towards the computational challenge. In *P5 eHealth: An Agenda for the Health Technologies of the Future*. Ed. by G. Pravettoni and S. Triberti. Springer International Publishing, 71–86. DOI: 10.1007/978-3-030-27994-3\_5.
- S. Van Buuren 2012. *Flexible imputation of missing data*. CRC Press. ISBN: 978-1-4398-6825-6.
- B. Van Calster, L. Wynants, D. Timmerman, E. W. Steyerberg, and G. S. Collins 2019. Predictive analytics in health care: how can we know it works? *Journal of the American Medical Informatics Association* 26 (12), 1651–1654. DOI: 10.1093/jamia/ocz130.
- S. Varma and R. Simon 2006. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics* 7, 91. DOI: 10.1186/1471-2105-7-91.
- A. Vellido 2020. The importance of interpretability and visualization in machine learning for applications in medicine and health care. *Neural Computing and Applications* 32 (24), 18069–18083. DOI: 10.1007/s00521-019-04051-w.
- V. Vesterinen, A. Nummela, S. Äyrämö, T. Laine, E. Hynynen, J. Mikkola, and K. Häkkinen 2016. Monitoring training adaptation with a submaximal running test under field conditions. *International Journal of Sports Physiology and Performance* 11 (3), 393–399. DOI: 10.1123/ijsp.2015-0366.
- M. Viceconti, P. Hunter, and R. Hose 2015. Big data, big knowledge: big data for personalized healthcare. *IEEE Journal of Biomedical and Health Informatics* 19 (4), 1209–1215. DOI: 10.1109/JBHI.2015.2406883.
- J. van der Waa, T. Schoonderwoerd, J. van Diggelen, and M. Neerincx 2020. Interpretable confidence measures for decision support systems. *International Journal of Human-Computer Studies* 144, 102493. DOI: 10.1016/j.ijhcs.2020.102493.
- J. Wainer and G. Cawley 2021. Nested cross-validation when selecting classifiers is overzealous for most practical applications. *Expert Systems with Applications* 182, 115222. DOI: 10.1016/j.eswa.2021.115222.
- S. A. Waldman and A. Terzic 2019. Healthcare evolves from reactive to proactive. *Clinical Pharmacology and Therapeutics* 105 (1), 10–13. DOI: 10.1002/cpt.1295.

- F. Walther, L. Heinrich, J. Schmitt, M. Eberlein-Gonska, and M. Roessler 2022. Prediction of inpatient pressure ulcers based on routine healthcare data using machine learning methodology. *Scientific Reports* 12, 5044. DOI: 10.1038/s41598-022-09050-x.
- WHO 2001. *International Classification of Functioning, Disability and Health: ICF*. World Health Organization, Geneva.
- W. Wolberg, O. Mangasarian, N. Street, and W. Street 1995. Breast cancer Wisconsin (diagnostic). UCI Machine Learning Repository. DOI: 10.24432/C5DW2B.
- T.-T. Wong 2015. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition* 48 (9), 2839–2846. DOI: 10.1016/j.patcog.2015.03.009.
- L. Wynants, M. Van Smeden, D. J. McLernon, D. Timmerman, E. W. Steyerberg, and B. Van Calster 2019. Three myths about risk thresholds for prediction models. *BMC Medicine* 17, 192. DOI: 10.1186/s12916-019-1425-3.
- Q. Xue, X. Wang, S. Meehan, J. Kuang, J. A. Gao, and M. C. Chuah 2018. Recurrent neural networks based obesity status prediction using activity data. In 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA), 865–870. DOI: 10.1109/ICMLA.2018.00139.
- L. Yang and A. Shami 2020. On hyperparameter optimization of machine learning algorithms: theory and practice. *Neurocomputing* 415, 295–316. DOI: 10.1016/j.neucom.2020.07.061.
- S.-I. Yi, B.-R. So, C.-S. Lee, S.-J. Lee, S.-K. Park, B.-K. Park, and I.-W. Chung 2011. Classification of health grade using bio-check unit and health index. *Journal of Biomechanical Science and Engineering* 6 (3), 148–159. DOI: 10.1299/jbse.6.148.
- M. J. Zaki, W. Meira Jr, and W. Meira 2014. *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press. ISBN: 978-0-521-76633-3.
- S. Zare, M. R. Thomsen, R. M. Nayga, and A. Goudie 2021. Use of machine learning to determine the information value of a BMI screening program. *American Journal of Preventive Medicine* 60 (3), 425–433. DOI: 10.1016/j.amepre.2020.10.016.
- J. Zou and L. Schiebinger 2018. AI can be sexist and racist – it’s time to make it fair. *Nature* 559, 324–326. DOI: 10.1038/d41586-018-05707-8.



## ORIGINAL PAPERS

### I

# **PREDICTING OVERWEIGHT AND OBESITY IN LATER LIFE FROM CHILDHOOD DATA: A REVIEW OF PREDICTIVE MODELING APPROACHES**

by

Ilkka Rautiainen and Sami Äyrämö 2021

Computational Sciences and Artificial Intelligence in Industry: New Digital  
Technologies for Solving Future Societal and Economical Challenges, pages  
203 – 220

[https://doi.org/10.1007/978-3-030-70787-3\\_14](https://doi.org/10.1007/978-3-030-70787-3_14)

Reproduced with kind permission of Springer International Publishing.



# Predicting Overweight and Obesity in Later Life from Childhood Data: A Review of Predictive Modeling Approaches



Ilkka Rautiainen and Sami Äyrämö

**Abstract** Overweight and obesity are an increasing phenomenon worldwide. Reliable and accurate prediction of future overweight or obesity early in the childhood could enable effective interventions by experts. While a lot of research has been done using explanatory modeling methods, capability of machine learning, and predictive modeling, in particular, remain mainly unexplored. In predictive modeling, the models are validated with previously unseen examples, giving a more accurate estimate of their performance and generalization ability in real-life scenarios. Our objective was to find and review existing overweight or obesity research from the perspectives of childhood data and predictive modeling. Thirteen research articles and three review articles were identified as relevant for this review. In general, prediction models with high performance either have a short time span to predict and/or are based on late childhood data. Logistic regression is currently the most often used method in forming the prediction models, although recently more complex models have also been applied. In addition to child's own weight and height information, maternal weight status and body mass index were often used as predictors in the models. More recent research has started to focus on a wider variety of other predictors as well.

**Keywords** Predictive models · Machine learning · Artificial intelligence · Obesity · Overweight

## 1 Introduction

Obesity is a global phenomenon that has increased rapidly during the last few decades in most countries. This trend has also led to significant increase in obesity-related diseases and deaths [15]. The review presented here aims to map and review the

---

I. Rautiainen (✉) · S. Äyrämö

Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, 40014 Jyväskylä, Finland

e-mail: [ilkka.t.rautiainen@jyu.fi](mailto:ilkka.t.rautiainen@jyu.fi)

S. Äyrämö

e-mail: [sami.ayramo@jyu.fi](mailto:sami.ayramo@jyu.fi)

© Springer Nature Switzerland AG 2022

T. Tuovinen et al. (eds.), *Computational Sciences and Artificial Intelligence in Industry, Intelligent Systems, Control and Automation: Science and Engineering 76*,

[https://doi.org/10.1007/978-3-030-70787-3\\_14](https://doi.org/10.1007/978-3-030-70787-3_14)

methods used in overweight and obesity prediction research with an emphasis on predictive modeling techniques. The basic question to consider is “Can we predict a person’s overweight or obesity status in later life from the data collected during childhood?” Ideally, early identification will make it possible to take steps for a successful obesity intervention. Currently, this identification is often done manually by using growth references such as de Onis et al. [12], Saari et al. [26], and Cole and Lobstein [9]. Also, the data collection phase can be a tedious and expensive process. It is therefore preferred that this identification can be achieved with easily available basic data, such as height and weight information. A widely used measure derived from these two attributes is body mass index (BMI). Ideally, children in unhealthy BMI trajectories should be identified before school age [21]. For adults, the BMI cut-off points in widest use are  $25 \text{ kg/m}^2$  for overweight and  $30 \text{ kg/m}^2$  for obesity [11]. Age and sex specific cut-off values that can be used in children have also been developed [10, 11].

When validating the employed prediction model, the performance measures are always a tradeoff between sensitivity (in this case the ratio of correctly classified overweight or obese cases in relation to total overweight or obese cases) and specificity (the ratio of correctly classified normal weight cases in relation to total normal weight cases). In this context, the most important performance measure is sensitivity, which indicates the proportion of children we are most interested in finding, in order to target preventive actions in this group [34]. However, when the specificity is low, Butler et al. [5] argue that it becomes questionable to use the model at all, since the model generates large amounts of false positives. They argue further that the risk threshold should be placed based on considering multiple criteria, including potential risks and harms as well as financial costs. Additionally, if a suitable tool for obesity prediction became eventually available for clinical use, the practitioners should take careful steps to avoid potential ethical issues in interventions. Communicating the overweight/obesity risk to child’s parents might have undesirable and unforeseen consequences. Also, availability of practical remedies should be ensured for those deemed to be at significant risk for obesity [20].

Breiman [4] and Shmueli [30] have discussed the differences between explanatory and predictive types of modeling. Explanatory modeling is defined as “the use of statistical models for testing causal explanations” and predictive modeling as “the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations” [30]. Validation is used in predictive modeling to estimate how well the model is expected to work with previously unseen data [3]. According to Bzdok et al. [6], prediction makes it possible to identify the best courses of action without requirement to understand the underlying mechanisms. In explanatory modeling, the model is not validated with an independent test data and thus its performance in the task of predicting new observations might be overestimated [28].

## 2 Materials and Methods

The scope of this review included studies in English language that

1. Predict the future overweight/obesity status with a model built using the baseline data collected in childhood (e.g., using the data collected up to the age of three years to predict child's obesity status at the age of six years). This criterion includes only studies that either build binary classifiers or somehow employ regression in solving the classifier problem.
2. In addition to describing the prediction framework also report relevant numeric results (e.g., sensitivity and specificity and/or AUC values).
3. Validate the results using an internal and/or external independent test set that has not been used for training the model. Internal bootstrap validation is not considered here to be an independent test data set.

Exceptions to these criteria were made for existing surveys, literature reviews, and meta-analyses directly concerned in a similar prediction problem. An overview of the study selection process is presented in Fig. 1. Studies fulfilling the criteria were searched initially from two databases, PubMed<sup>1</sup> and IEEE Xplore,<sup>2</sup> in addition to extensive searches on Google Scholar.<sup>3</sup> In the initial search phase, potential studies were mapped based on the titles of studies. PubMed search term was

```
((bmi[Title] OR (body mass index[Title])) AND (obesity[Title] OR obese[Title] OR overweight[Title]) AND (prediction[Title/Abstract] OR predicting[Title/Abstract])).
```

This search yielded 63 results in total, of which four were identified as potential for inclusion. For IEEE Xplore, “obesity” search term was used with conference articles, journals, and magazines starting from 1994 included in the search. The search yielded 634 results with two potential studies. An additional set of potential studies were then searched using Google Scholar, by using variations of the search terms described before. These initial searches were made in July 2017. After this initial seed set of potential studies was identified, references in studies and newer research that cite the studies were mapped for finding additional potentially relevant studies. Some of the relevant studies were identified already during these steps.

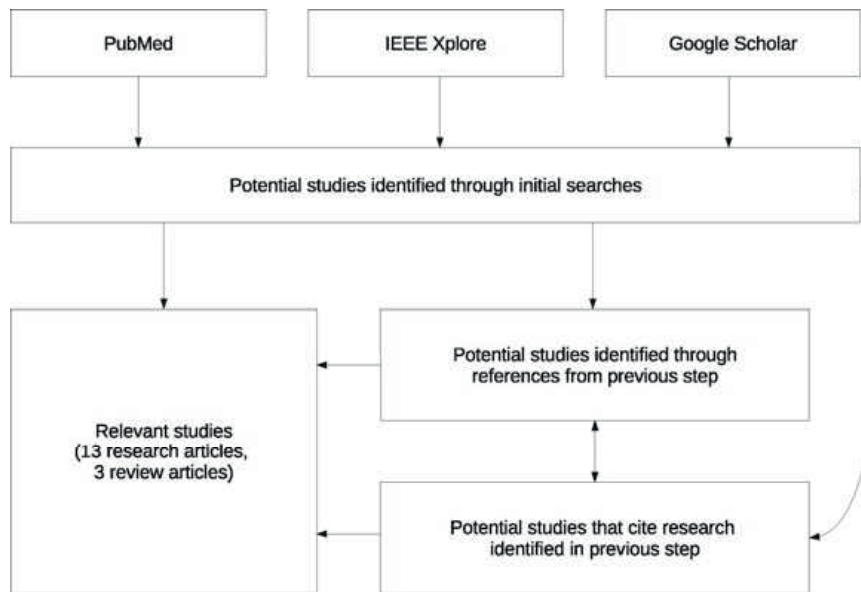
A second search phase is indicated by a two-way arrow in Fig. 1. This searching through references of the potential studies and research that cited the potential studies continued until no new potential studies emerged. Each search cycle produced new material to the pool of potentially relevant studies. The idea and assumption behind

---

<sup>1</sup> <https://www.ncbi.nlm.nih.gov>.

<sup>2</sup> <http://ieeexplore.ieee.org>.

<sup>3</sup> <https://scholar.google.com>.



**Fig. 1** An overview of the study selection process

this approach was that the relevant studies are most likely referenced in recent (2010–2018) research articles. Hence the search focused on the references of recent studies that were identified as potential for inclusion. This search phase lasted until August 2018. To find the most recent studies released after 2018, the second search phase was repeated in July 2020. During the final search, five new articles released at the latest in June 2020 were identified and added to the set of included studies.

### 3 Results

To our knowledge, this is the first study collecting the existing research that focuses on the predictive modeling approach in prediction of overweight or obesity. Three surveys, literature reviews and meta-analyses from the years 2010 to 2015 closely related to the problem of overweight/obesity prediction are presented and discussed in Sect. 3.1.

Eight out of 13 studies from 2004 to 2020 presented in Sect. 3.2 employ logistic regression in their models. Only two early studies [14, 34] explore more complex models such as decision trees, Bayesian and neural networks, support vector machines, that have greater capacity to learn associations from data. Recently, in 2019 and 2020, more complex models [17, 18, 24, 32] have been explored more extensively. Seven studies predict only obesity and five only overweight including obesity. One study [7] predicts separately two cases: overweight (including obesity) and obesity. The age range to predict overweight/obesity status in the studies was from two years to 33 years, with ten studies predicting status in children under ten

years of age. In addition to weight and height information of the child, number of features used in prediction varied from zero to over 1700. Generally, more complex models employed recently used more features than simpler models presented in the earlier studies.

### ***3.1 Surveys, Literature Reviews, and Meta-analyses***

A survey of data mining methods used in the field was conducted by Adnan et al. [1]. The study summarized some of the research in the area and focused on describing three methods used in three different studies: neural networks, naïve Bayes classifiers, and decision trees. Each of the methods was described as having its own strengths and weaknesses. The conclusion was that to improve prediction results, further improvements of the techniques are necessary. The authors planned to continue their work by combining different existing methods to form a single better performing hybrid method.

Infancy weight gain was used as a predictor for childhood obesity in a meta-analysis presented by Druet et al. [13]. The meta-analysis reported a consistent positive association of infancy weight gain to subsequent obesity. The study included ten cohort studies from which full data of three cohorts ( $N = 8236$ ) were used in forming models for overweight and obesity prediction. A second sample of the same size was used for validating the results. The prediction model employed stepwise multivariable logistic regression, and used mother's BMI, child's birthweight and sex in addition to weight gain information from birth to one year of age for childhood obesity prediction. The model was reported to show moderate predictive ability with area under curve (AUC) value of 0.77. When using a risk score threshold that puts 10% of the population above the threshold, the model had a sensitivity of 58.6% and specificity of 90.9%. A similar model was also created for childhood overweight prediction and reported AUC value was 0.76.

An extensive systematic review and meta-analysis of existing studies were presented by Simmonds et al. [31] to examine the use of different measures of obesity in childhood for predicting obesity and development of obesity-related diseases. The study found that the predictive accuracy of childhood obesity for predicting adult obesity had a sensitivity of 30% and a specificity of 98% and was described as moderate. The study concluded that childhood BMI is not an effective predictor for obesity or disease in adulthood, since most obese adults were not obese in childhood. However, no evidence was found in the study to support any other single measure over BMI.

### 3.2 Predictive Modeling Approaches

An overview of the studies employing predictive modeling is presented in Table 1. The studies might present also models that have not been validated with an independent data set, i.e., be partly explanatory in their approach. *Predicted status* (overweight/obesity) as well as other information is listed only if the results were validated independently. The number of *additional features* used exclude sex, weight, height, BMI of the child and overweight/obesity labels. If the study had separate models for females and males, best results (AUC, sensitivity, and specificity) are presented for both. Two of the studies [16, 34] presented different models for predicting overweight based on data collected up to three points of time. The best result for each of these models is also presented in the tables. In addition to these, Cheung et al. [7] presented separate models for obesity and overweight prediction. Again, the best results for both models are presented.

Data collected at childhood was used by Cheung et al. [7] to predict obesity and overweight at the age of 33 years and self-reported disease history at the age of 42 years. ROC analysis was used to define optimal BMI risk thresholds at the ages of seven, 11, and 16 years. The cohort data was split in half, with the first half used for the ROC analysis and the second one for validating the cut-off BMI thresholds. For the validation data ( $N = 4231$ ) at the age of 11 years, sensitivity and specificity values for obesity status prediction were, respectively, 71.7% and 72.4% for males. For overweight prediction, the same values were 65.6% and 68.6%. Obesity prediction for females achieved sensitivity of 75.7% and specificity of 69.7%, while the same values for overweight prediction were 69.8% and 63.6%.

Six independent data mining methods were investigated and compared with the conventional logistic regression approach by Zhang et al. [34]. The study included C4.5 decision tree, Bayesian networks, naïve Bayes, association rules, neural networks, and support vector machines (SVMs). The task was to predict overweight status at the age of three years based on child's data ( $N = 16523$ ) collected up to six weeks, eight months, and two years, respectively. 67% of the data was used for training the models and 33% for testing. Children's height and weight data were used in the models. Time of gestation was used as an additional feature when predicting overweight status from the age of two years. The results showed that the prediction accuracy improved with the methods used when compared with the logistic regression approach. The best performing algorithms were SVM with radial basis function (RBF) kernel and Bayesian methods, specifically naïve Bayes. At the age of two years, the sensitivity and specificity of NB were 54.7% and 93.1%, respectively, while the same values for SVM were 60.0% and 79.6%. The conclusion was that to improve overweight prediction rates more features may need to be recorded and used for prediction.

The aim of Dugan et al. [14] was to improve the work presented in Zhang et al. [34] by considering a significantly extended set of predictors. A partly different set of machine learning algorithms were also explored in the study. They included random tree, random forest, ID3, and C4.5 decision trees in addition to naïve Bayes and



**Table 1** Overview of the studies using predictive modeling methods for overweight/obesity prediction based on childhood data

Study	Method(s)	Predicted status	Age(s) to predict	Predicts status from child's data recorded up to	Weight and height information used	Number of additional features used	Additional features	Validation AUC (best)	Validation sensitivity (best)	Validation specificity (best)	N (training data)	N (validation/testing data)
Cheung et al. [7]	ROC threshold cut-off	Overweight (incl. obesity) and obesity	33 years	11 years	BMI	0	N/a	N/a	75.7%*, 71.7%** (obesity), 69.8%*, 65.6%** (overweight)	69.7%*, 72.4%** (obesity), 63.6%*, 68.6%** (overweight)	4136	4231
Zhang et al. [34]	Logistic regression, decision tree, association rules, neural network, linear SVM, RBF SVM, Bayesian network, naïve Bayes	Overweight (incl. obesity)	3 years	6 weeks, 8 months, 2 years	Birth weight, weight change, height, BMI	0 (6 weeks), 0 (8 months), 1 (2 years)	Time of gestation	N/a	11.2% (6 weeks), 35.5% (8 months), 54.7% (2 years)	96.0% (6 weeks), 91.5% (8 months), 93.1% (2 years)	≈11070 <sup>g</sup>	≈5400–5600 <sup>h</sup>
Dugan et al. [14]	Random tree, random forest, ID3 decision tree, J48 decision tree, naïve Bayes, Bayesian network	Obesity	2–10 years <sup>a</sup>	<2 years	Height and weight	Over 160	Multiple <sup>d</sup>	N/a	89%	83%	≈6767 per fold (10-fold cv)	≈752 per fold (10-fold cv)

(continued)

**Table 1** (continued)

Study	Method(s)	Predicted status	Age(s) to predict	Predicts status from child's data recorded up to	Weight and height information used	Number of additional features used	Additional features	Validation AUC (best)	Validation sensitivity (best)	Validation specificity (best)	N (training data)	N (validation/testing data)
Morandi et al. [22]	Stepwise logistic regression	Obesity	7 years	Birth	Birth weight	5	Maternal BMI, paternal BMI, number of household members, maternal occupation, gestational smoking	0.73	N/a	N/a	4032	1503 and 1032
Santorelli et al. [29]	Stepwise logistic regression	Obesity	2 years	9 months, 12 months	Birth weight and weight change	0	N/a	0.850 (9 months), 0.886 (12 months)	N/a	N/a	1528 (9 months), 731 (12 months)	880 (9 months) 867 (12 months)
Weng et al. [33]	Stepwise logistic regression	Overweight (incl. obesity)	3 years	1 year	Birth weight and weight change	4	Maternal pre-pregnancy weight status, paternal BMI, maternal smoking during pregnancy, breastfeeding in the first year	0.755	76.9%	66.5%	8299	1715

(continued)



**Table 1** (continued)

Study	Method(s)	Predicted status	Age(s) to predict	Predicts status from child's data recorded up to	Weight and height information used	Number of additional features used	Additional features	Validation AUC (best)	Validation sensitivity (best)	Validation specificity (best)	N (training data)	N (validation/testing data)
Redsell et al. [25]	Stepwise logistic regression	Overweight (incl. obesity)	5 years	1 year	Birth weight and weight change	4	Maternal pre-pregnancy weight status, paternal BMI, maternal smoking during pregnancy, breastfeeding in the first year	0.79	53%	71%	8299	980
Graversen et al. [16]	Logistic regression	Overweight (incl. obesity)	13–16 years <sup>b</sup>	Birth, 5 years, 8 years	Birth weight and BMI	1	Maternal BMI	N/a	24.0%*, 17.4%** (birth), 38.9%*, 28.2%** (5 years), 49.2%*, 38.7%** (8 years)	92.1%*, 91.7%** (birth), 94.4%*, 94.2%** (5 years), 96.0%*, 96.7%** (8 years)	4111	5414
Hammond et al. [18]	Logistic regression, LASSO regression, random forest classifier and regression, gradient boosting classifier and regression	Obesity	5 years	2 years	Weight-for-length and BMI	1509 in total <sup>c</sup>	Multiple <sup>e</sup>	0.817*, 0.761**	0.491* <sup>f</sup> , 0.699** <sup>f</sup>	0.912* <sup>f</sup> , 0.670** <sup>f</sup>	2760	689

(continued)

**Table 1** (continued)

Study	Method(s)	Predicted status	Age(s) to predict	Predicts status from child's data recorded up to	Weight and height information used	Number of additional features used	Additional features	Validation AUC (best)	Validation sensitivity (best)	Validation specificity (best)	N (training data)	N (validation/testing data)
Saijo et al. [27]	Logistic regression	Obesity	6–8 years	1 year	Obesity index at one year of age	3	Information about the mother: pre-pregnancy BMI, smoking during pregnancy, education	0.688	49.7%	83.8%	5488	1358
Pang et al. [24]	XGBoost	Obesity	2–7 years	2 years	Weight, height, BMI-for-age and other derived information	102 in total	Multiple features incl. demographic features and laboratory values	0.81	79.91%	63.27%	≈21762	≈5441
Gupta et al. [17]	Recurrent neural network (RNN) with LSTM	Obesity	All ages from 3 up to 20 years	Multiple observation windows employ data from age of 0 to 17 years	At least weight, height, BMI and other values generated from the data	1737 in total	Thousands of variables derived from various EHR data	One-year: 0.87–0.97, two-year: 0.82–0.96, three-year: 0.80–0.95 (estimated from the plots)	One-year: 0.63–0.91, two-year: 0.66–0.92, three-year: 0.63–0.87	N/a	≈40817	≈13606

(continued)

**Table 1** (continued)

Study	Method(s)	Predicted status	Age(s) to predict	Predicts status from child's data recorded up to	Weight and height information used	Number of additional features used	Additional features	Validation AUC (best)	Validation sensitivity (best)	Validation specificity (best)	N (training data)	N (validation/testing data)
Singh and Tawfik [32]	kNN, J48 pruned tree, random forest bagging, SVM, multilayer perceptron (MLP), voting	Overweight (incl. obesity)	14 years	11 years	BMI	0	N/a	N/a	62–94% (depending on the setting)	N/a	≈7777	≈3333

\*Females

\*\*Males

<sup>a</sup>If the child was obese at any time between the ages of two and ten years, he/she was labeled as obese

<sup>b</sup>If the child was overweight at any time between the ages of 13 and 16 years, he/she was labeled as overweight

<sup>c</sup>Original feature space of 19290 features shrunk to 8% when a minimum of five children with information was used as a threshold

<sup>d</sup>167 variables in total, but they are not listed in detail. A questionnaire form used to collect the information is presented

<sup>e</sup>Available features are listed using 23 main categories

<sup>f</sup>The best model was presented as a performance tradeoff table. The values presented here are taken from the point of the highest F1 score value

<sup>g</sup>67% of total data

<sup>h</sup>33% of total data from 16523 equals 5453 observations. However, the confusion matrices in Figs. 3 and 4 in the study include 5618 observations

Bayesian networks. The study employed in total 167 predictors that were collected through questionnaires filled by parents and physicians. The data ( $N = 7519$ ) were collected before the children's second birthday. 10-fold cross-validation was used for validating the results. If the child was obese at any point after her/his second birthday to the age of 10 years, the child was labeled as obese. The prediction task consisted of predicting the child's future obesity status after the second birthday. Two of the algorithms, ID3 and NB, performed slightly better when predictors described as "noisy" were removed from the data. The ID3 model used 87 and the NB model 107 out of 167 predictors. Other methods employed in the study performed best when all the available predictors were utilized. The study singled out two algorithms as best working with this data set: ID3 decision tree with sensitivity and specificity of 89% and 83%, respectively, along with random tree's performance metrics of 88% and 80%.

Stepwise logistic regression analysis was used to form predictive models by Morandi et al. [22] to predict obesity status at the age of seven years with data available at birth. The final model included maternal and paternal BMI, number of household members, maternal occupation, gestational smoking as well as birth weight. First cohort data ( $N = 4032$ ) was used in training the model, while validation of the results was performed using two independent cohorts ( $N_1 = 1503$ ,  $N_2 = 1032$ ). Reported AUC values for the two validation sets were 0.70 and 0.73 respectively. For the first validation cohort, gestational smoking and number of household members information had to be omitted from the model. The study also experimented with adding a genetic score information to the models, but that did not provide any significant improvement in terms of predictive power.

Three logistic regression models were formed with stepwise method by Santorelli et al. [29], where the models were applied in a mobile application aimed at parents. The models predicted the risk of childhood obesity at the age of two years. The plan presented in the study was that the parents use the application when their infant's age reaches six, nine, and twelve months, with separate models presented for each case. Input features included sex, birth weight, and weight gain. Two datasets were used. The first one was used to form the three models ( $N_1 = 1022$ ,  $N_2 = 1528$ ,  $N_3 = 731$ ) and validated using an internal bootstrap validation. The second dataset was used as a separate external testing data. The second and third models were validated using this data ( $N_2 = 880$ ,  $N_3 = 867$ ), with the reported AUC values for the two models being 0.85 and 0.89, respectively. More detailed sensitivity and specificity values were also reported for different model configurations, but only for the training data. The mobile application has since been discontinued, and no published research exists on the usage or effectiveness of the application [5].

Weng et al. [33] formed a childhood overweight prediction model (IROC) using stepwise logistic regression for predicting obesity status at the age of three years. The cohort data was randomly divided into training ( $N = 8299$ ) and testing ( $N = 1715$ ) sets. The data included 33 potential predictor variables. From these, seven significant input features were identified: sex, birth weight, weight gain in first year, maternal pre-pregnancy weight status, paternal BMI, maternal smoking during pregnancy and breastfeeding in the first year. The study reported a moderately good predictive

ability, with sensitivity value of 0.769, specificity of 0.665 and AUC value 0.755 for the testing data.

The IROC algorithm [33] was further validated by Redsell et al. [25] with an additional independent dataset ( $N = 980$ ), predicting overweight status at the age of five years. Four models were formed for prediction. The first one (clinical model) used the original algorithm directly and assigned null values to missing data. The reported AUC values were 0.67 when using the International obesity taskforce overweight criteria [11] and 0.65 when using the UK 1990 overweight criteria [10]. The second one (recalibrated model) used multivariate logistic regression to generate and recalibrate the model to reflect the demographics of the new validation data (AUC values of 0.70 and 0.67 were reported). The third one (imputed model) used multiple imputation to generate ten copies of the existing data set to predict missing risk factor from multivariate models (AUC values 0.79 and 0.73). The fourth one (recalibrated imputed model) applied the recalibrated algorithm to the imputed data (AUC values 0.93 and 0.90).

Graversen et al. [16] developed logistic regression models to predict adolescent overweight, adult overweight, and adult obesity. Input features for the models included maternal BMI, birth weight, and early childhood BMI. First dataset ( $N = 4111$ ) was used to form the models. Performance of the models was validated with an internal bootstrap validation as well as with an external independent dataset ( $N = 5414$ ), where the prevalence of overweight was much higher. Only the model for adolescent overweight prediction was validated with the external data. The study reported results for adolescent overweight prediction from data collected at birth, up to the age of five years, and up to the age of eight years. Also, different thresholds of percentage of children labeled as “at risk” of overweight were explored. With the external validation dataset, sensitivity and specificity for adolescent overweight prediction at the age of five years for females were 38.9% and 94.4%, respectively, when the threshold of being at risk was set to upper 10%. For males these measures were 28.2% and 94.2%.

Hammond et al. [18] employed a set of classifier and regression models for predicting the obesity status at the age of five years. The regression models were configured to predict the BMI value, which was then converted to a binary value indicating if the child was predicted to be obese or not. The models were formed separately for girls and boys. Their initial set ( $N = 3449$ ) of predictors was very large at 19,290 features, but shrunk down to 8% when a minimum of five children with information was used as a threshold. For validation, the study held out a random test set that included 20% of the data. The remaining data was then used for training the model with an internal bootstrap validation. The best performing models with the highest mean test data AUC values (0.817 for girls and 0.761 for boys) were achieved with LASSO regression for both sexes. The final LASSO regression model included 36 features for girls and 50 features for boys. The features associated strongest with obesity for both sexes were weight-for-length z-score and BMI.

Saijo et al. [27] divided a total sample of mother-child pairs ( $N = 6846$ ) randomly into derivation (80%) and validation (20%) cohorts. The study identified a set of obesity predictors based on univariate and multivariate logistic regression analysis

employing data from pregnant women and one year old infants. The predictors identified in the model derivation phase included mother's pre-pregnancy BMI, child's gender, smoking during pregnancy, education, and obesity index at the age of one year. The predicted outcome was the obesity index of over 20% from the weight-for-height charts for Japanese children at the age of 6–8 years. The reported AUC for the validation cohort was 0.688, with sensitivity of 68.8% and specificity of 83.8%.

Pang et al. [24] employed the gradient boosting library XGBoost for their obesity prediction model. The target was to predict, based on the child's data collected up to the age of two years, if the child was to be obese at any stage between the ages of two and seven years. The data ( $N = 27203$ ) was divided into two separate training sets (40% each) and a test set (20%). The training sets were utilized in hyperparameter optimization conducted using 10-fold cross-validation. The study highlighted two new predictive features related to obesity: body temperature and respiratory rate. Other predictive features reported were weight, height, race, and ethnicity. The model AUC for the test set at sensitivity level of 80% was 0.81 with specificity of 63.27%.

Gupta et al. [17] approached the problem of obesity prediction with deep learning, using recurrent neural network (RNN) architecture with long short-term memory (LSTM) cells. The models were configured to predict the BMI value. The data ( $N = 68029$ ) consisted of over 44 million records. After cleaning the variables that did not have enough information, the final data consisted of 1737 variables in total. Thousands of variables available in the EHR system are employed to form the time-series data for the patients. The time-series data included thousands of variables derived from condition diagnosis, drug prescription data, information on performed procedures, and laboratory results. In addition, there were static variables indicating sex, race, ethnicity, and zip code. 10-fold cross-validation was employed, with 60% of the data used for training, 20% for validation and 20% for the final test set. Data from 16 two-year observation windows were used to predict obesity at prediction windows placed at one, two, and three years ahead of the observation window. In total, the prediction models were formed for 48 sub-cohorts. The prediction results varied between sub-cohorts. For one-year predictions reported AUC was between 0.87–0.97, for two-year predictions 0.82–0.96, and for three-year predictions 0.80–0.95 (when estimated from the reported plots). Additional measures reported included sensitivity, accuracy, and positive predictive value (e.g., precision). The three most important variables obtained by averaging the importance scores were BMI (percentile) per age and gender, obese/non-obese label, and allergic urticaria. Their work also aimed to add interpretability to the proposed model. Interpretability of the results was also considered on two levels: time-level and variable-level.

Singh and Tawfik [32] compared seven different classifier methods in the overweight prediction domain. For each child in the study ( $N = 11110$ ), BMI was recorded at the ages of 3, 5, 7, and 11 years. This information was used in the prediction of overweight status at the age of 14 years. The results were reported for two cases. In the first case, the models were trained using the original imbalanced data. The best-reported sensitivity, attained using the J48 pruned tree, in the “normal” class was 0.94, with “at risk” class sensitivity of 0.62. Corresponding precision values were 0.87 and 0.78. In the second case, an oversampling method was employed

to the minority class to balance the data. Training the models with the balanced data improved the sensitivity of the “at risk” class. The sensitivity for the “normal” class in the second case with the MLP method was 0.62, with “at risk” class sensitivity of 0.92. The corresponding precision values were 0.96 and 0.51.

## 4 Conclusions

This review explored the existing research on overweight and obesity prediction. While various explanatory models have been studied and employed extensively in the research area, utilization of predictive modeling methods of machine learning remains partly unexplored in the field.

The studies using explanatory modeling do not validate the formed models with a separate test data. Instead they examine how well the whole data fits to the model. Generalization refers to how well the model trained on the training data set predicts the output for new instances, and it is an integral part of the machine learning and predictive modeling approach [2, 3]. We argue that if the model is aimed for prediction, the model should always be validated with independent data to get more reliable performance estimates.

In terms of predictive power, best performing models in the study either made the prediction quite late or had a relatively short period between the prediction and the outcome. Graverson et al. [16] had a very high specificity (96%) in predicting overweight at adolescence in girls, using data recorded up to eight years. Other moderately successful models [14, 24, 29] and “two years to three years” model by Zhang et al. [34] only had a short time period to cover. Recent studies [17, 18], which have employed more data and models with more complexity, have shown some promise when making predictions over longer time periods. So far there has been no evidence of success in employing any of the presented or other models in clinical use [5].

The constantly growing size of data sets will enable the use of even more powerful machine learning methods and, moreover, sophisticated analysis of the obtained models. In the present context, one of the most promising approach is, for example, the recurrent neural networks [19], currently explored only in one study [17]. They could provide a powerful method for predicting overweight and obesity development in the later life, because they are inherently designed for time-series prediction tasks (e.g., [8]). It is, however, important to collect a sufficient amount of data before training complex models, such as neural networks.

Besides being powerful methodology for building prediction models, more effort could also be directed to machine learning-based hypothesis generation by employing large data sets, high-performance computing and machine learning algorithms. This approach can lead to finding of unsuspected information and predictors from the growing data.

A critical issue to be considered when applying machine learning in overweight and obesity prediction tasks is, however, the risk of chance findings. In order to



minimize the risk of chance models or predictors, it is highly important to develop and apply strategies, such as data randomization [23], for confirming significance of the obtained models and relevance of the identified predictors.

## 5 Summary Points

What was already known on the topic

- A lot of research has been done on predicting upcoming obesity or overweight.
- None of the prediction models presented in previous research have been successfully utilized in practice.

What this study added to our knowledge

- Although prediction in domain of overweight/obesity has been studied extensively, there are not many studies that separate the training and validation of the model in a way expected in predictive modeling.
- Highest performing prediction models are either only predicting near future overweight/obesity status or make their prediction relatively late.
- More complex models with greater learning capacity have been employed in only a limited set of studies. Future research possibilities in the area exist in examining the potential of recurrent models further and considering further model building practices such as predictor selection and significance of prediction.

**Acknowledgements** We thank Richard Allmendinger for his comments on the initial draft article. Ilkka Rautiainen received funding from Business Finland in addition to a grant from the Jenny and Antti Wihuri Fund. Sami Äyrämö also received funding from Business Finland. The funding sources did not have any other involvement in the study.

## References

1. Adnan MHB, Husain W, Damanhoori F (2010) A survey on utilization of data mining for childhood obesity prediction. In: 8th Asia-Pacific symposium on information and telecommunication technologies (Kuching, 2010). IEEE, pp 1–6
2. Alpaydin E (2014) Introduction to machine learning, 3rd edn. MIT Press, Cambridge
3. Bishop C (2006) Pattern recognition and machine learning. Springer, Berlin
4. Breiman L (2001) Statistical modeling: the two cultures. *Stat Sci* 16(3):199–215. <https://doi.org/10.1214/ss/1009213726>
5. Butler ÉM, Derraik JGB, Taylor RW, Cutfield WS (2018) Childhood obesity: how long should we wait to predict weight? *J Pediatr Endocrinol Metab* 31(5):497–501. <https://doi.org/10.1515/jpem-2018-0110>
6. Bzdok D, Altman N, Krzywinski M (2018) Statistics versus machine learning. *Nat Methods* 15(4):233–234. <https://doi.org/10.1038/nmeth.4642>
7. Cheung YB, Machin D, Karlberg J, Khoo KS (2004) A longitudinal study of pediatric body mass index values predicted health in middle age. *J Clin Epidemiol* 57(12):1316–1322. <https://doi.org/10.1016/j.jclinepi.2004.04.010>



8. Choi E, Schuetz A, Stewart WF, Sun J (2017) Using recurrent neural network models for early detection of heart failure onset. *J Am Med Inform Assoc* 24(2):361–370. <https://doi.org/10.1093/jamia/ocw112>
9. Cole TJ, Lobstein T (2012) Extended international (IOTF) body mass index cut-offs for thinness, overweight and obesity. *Pediatr Obes* 7(4):284–294. <https://doi.org/10.1111/j.2047-6310.2012.00064.x>
10. Cole TJ, Freeman JV, Preece MA (1995) Body mass index reference curves for the UK, 1990. *Arch Dis Child* 73(1):25–29. <https://doi.org/10.1136/adc.73.1.25>
11. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* 320(7244):1240–1243. <https://doi.org/10.1136/bmj.320.7244.1240>
12. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ* 85(9):660–667. <https://doi.org/10.1590/S0042-96862007000900010>
13. Druet C, Stettler N, Sharp S, Simmons RK, Cooper C, Davey Smith G, Ekelund U, Lévy-Marchal C, Järvelin M-R, Kuh D, Ong KK (2012) Prediction of childhood obesity by infancy weight gain: an individual-level meta-analysis. *Paediatr Perinat Epidemiol* 26(1): 19–26. <https://doi.org/10.1111/j.1365-3016.2011.01213.x>
14. Dugan TM, Mukhopadhyay S, Carroll A, Downs S (2015) Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform* 6(3):506–520. <https://doi.org/10.4338/ACI-2015-03-RA-0036>
15. GBD (2017) Health effects of overweight and obesity in 195 countries over 25 years. *N Engl J Med* 377(1):13–27. (Written by GBD 2015 Obesity Collaborators). <https://doi.org/10.1056/NEJMoa1614362>
16. Gravensen L, Sørensen TIA, Gerds TA, Petersen L, Sovio U, Kaakinen M, Sandbaek A, Laitinen J, Taanila A, Pouta A, Järvelin M-R, Obel C (2015) Prediction of adolescent and adult adiposity outcomes from early life anthropometrics. *Obesity* 23(1):162–169. <https://doi.org/10.1002/oby.20921>
17. Gupta M, Phan T-LT, Bunnell T, Beheshti R (2020) Obesity prediction with EHR data: a deep learning approach with interpretable elements. [arXiv:1912.02655v5](https://arxiv.org/abs/1912.02655v5)
18. Hammond R, Athanasiadou R, Curado S, Aphinyanaphongs Y, Abrams C, Messito MJ, Gross R, Katzow M, Jay M, Razavian N, Elbel B (2019) Predicting childhood obesity using electronic health records and publicly available data. *PLoS One* 14(4). <https://doi.org/10.1371/journal.pone.0215571>
19. Hochreiter S, Schmidhuber J (1997) Long short-term memory. *Neural Comput.* 9(8):1735–1780. <https://doi.org/10.1162/neco.1997.9.8.1735>
20. Levine RS, Dahly DL, Rudolf MCJ (2012) Identifying infants at risk of becoming obese: can we and should we? *Public Health* 126(2):123–128. <https://doi.org/10.1016/j.puhe.2011.10.008>
21. Lynch BA, Rutten LJJ, Ebbert JO, Kumar S, Yawn BP, Jacobson D, Sauver JS (2017) Development of distinct body mass index trajectories among children before age 5 years: a population-based study. *J Prim Care Community Health* 8(4):278–284. <https://doi.org/10.1177/2150131917704326>
22. Morandi A, Meyre D, Lobbens S, Kleinman K, Kaakinen M, Rifas-Shiman SL, Vatin V, Gaget S, Pouta A, Hartikainen A-L, Laitinen J, Ruokonen A, Das S, Khan AA, Elliott P, Maffei C, Gillman MW, Järvelin M-R, Froguel P (2012) Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts. *PLoS One* 7(11). <https://doi.org/10.1371/journal.pone.0049919>
23. Ojala M, Garriga GC (2010) Permutation tests for studying classifier performance. *J Mach Learn Res* 11:1833–1863
24. Pang X, Forrest CB, Lê-Scherban F, Masino AJ (2019) Understanding early childhood obesity via interpretation of machine learning model predictions. In: 2019 18th IEEE international conference on machine learning and applications (ICMLA) (Boca Raton, FL, 2019). IEEE, pp 1438–1443. <https://doi.org/10.1109/ICMLA.2019.00235>

25. Redsell SA, Weng S, Swift JA, Nathan D, Glazebrook C (2016) Validation, optimal threshold determination, and clinical utility of the infant risk of overweight checklist for early prevention of child overweight. *Child Obes* 12(3):202–209. <https://doi.org/10.1089/chi.2015.0246>
26. Saari A, Sankilampi U, Hannila M-L, Kiviniemi V, Kesseli K, Dunkel L (2011) New Finnish growth references for children and adolescents aged 0 to 20 years: Length/height-for-age, weight-for-length/height, and body mass index-for-age. *Ann Med* 43(3):235–248. <https://doi.org/10.3109/07853890.2010.515603>
27. Saijo Y, Ito Y, Yoshioka E, Sato Y, Minatoya M, Araki A, Miyashita C, Kishi R (2019) Identifying a risk score for childhood obesity based on predictors identified in pregnant women and 1-year-old infants: an analysis of the data of the Hokkaido study on environment and children's health. *Clin Pediatr Endocrinol* 28(3):81–89. <https://doi.org/10.1297/cpe.28.81>
28. Sainani KL (2014) Explanatory versus predictive modeling. *PM R* 6(9):841–844. <https://doi.org/10.1016/j.pmrj.2014.08.941>
29. Santorelli G, Petherick ES, Wright J, Wilson B, Samiei H, Cameron N, Johnson W (2013) Developing prediction equations and a mobile phone application to identify infants at risk of obesity. *PLoS One* 8(8). <https://doi.org/10.1371/journal.pone.0071183>
30. Shmueli G (2010) To explain or to predict? *Stat Sci* 25(3):289–310. <https://doi.org/10.1214/10-STS330>
31. Simmonds M, Burch J, Llewellyn A, Griffiths C, Yang H, Owen C, Duffy S, Woolacott N (2015) The use of measures of obesity in childhood for predicting obesity and the development of obesity-related diseases in adulthood: a systematic review and meta-analysis. *Health Technol Assess* 19(43):1–336. <https://doi.org/10.3310/hta19430>
32. Singh B, Tawfik H (2020) Machine learning approach for the early prediction of the risk of overweight and obesity in young people. In: 20th international conference on computational science – ICCS 2020 (Amsterdam, 2020), Proceedings, Part IV. Springer, pp 523–535
33. Weng SF, Redsell SA, Nathan D, Swift JA, Yang M, Glazebrook C (2013) Estimating overweight risk in childhood from predictors during infancy. *Pediatrics* 132(2):e414–e421. <https://doi.org/10.1542/peds.2012-3858>
34. Zhang S, Tjortjis C, Zeng X, Qiao H, Buchan I, Keane J (2009) Comparing data mining methods with logistic regression in childhood obesity prediction. *Inf Syst Front* 11(4):449–460. <https://doi.org/10.1007/s10796-009-9157-0>



## II

### **PREDICTING FUTURE OVERWEIGHT AND OBESITY FROM CHILDHOOD GROWTH DATA: A CASE STUDY**

by

Ilkka Rautiainen, Jukka-Pekka Kauppi, Toni Ruohonen, Eero Karhu, Keijo  
Lukkarinen, and Sami Äyrämö 2021

Computational Sciences and Artificial Intelligence in Industry: New Digital  
Technologies for Solving Future Societal and Economical Challenges, pages  
189 – 201

[https://doi.org/10.1007/978-3-030-70787-3\\_13](https://doi.org/10.1007/978-3-030-70787-3_13)

Reproduced with kind permission of Springer International Publishing.

# Predicting Future Overweight and Obesity from Childhood Growth Data: A Case Study



**Ilkka Rautiainen, Jukka-Pekka Kauppi, Toni Ruohonen, Eero Karhu, Keijo Lukkarinen, and Sami Äyrämö**

**Abstract** Overweight, obesity and diseases associated with them have been increasing rapidly during the last few decades. The early detection of obesity risk is the key to preventive actions. This task can be supported by machine learning-based prediction models that reliably predict future overweight/obesity status based on early childhood data. The case study presented here employs predictive modeling and height/weight data collected by the health care system of Äänekoski town, Finland. The study utilizes nine existing study designs carefully selected based on a recent literature review. For each individual in the data, the BMI growth curves were resampled at 30-d intervals using the linear interpolation technique. This time series data is then utilized to form several predictive models using logistic regression, support vector machine, and a decision tree. Prediction accuracy is comparable to existing studies, and in some cases, even better. The best model, trained by the SVM method on the Finnish data, obtained an F1-score of 0.73. The results suggest that the Finnish data may contain strong dependencies that can be utilized in building the models. However, more versatile information from the early years of childhood is most likely needed to further optimize the models.

---

I. Rautiainen (✉) · J.-P. Kauppi · T. Ruohonen · S. Äyrämö  
Faculty of Information Technology, University of Jyväskylä, P.O. Box 35, FI-40014 Jyväskylä, Finland  
e-mail: [ilkka.t.rautiainen@jyu.fi](mailto:ilkka.t.rautiainen@jyu.fi)

T. Ruohonen  
e-mail: [toni.ruohonen@jyu.fi](mailto:toni.ruohonen@jyu.fi)

S. Äyrämö  
e-mail: [sami.ayramo@jyu.fi](mailto:sami.ayramo@jyu.fi)

E. Karhu · K. Lukkarinen  
Town of Äänekoski, Hallintokatu 4, FI-44100 Äänekoski, Finland  
e-mail: [eero.karhu@aanekoski.fi](mailto:eero.karhu@aanekoski.fi)

K. Lukkarinen  
e-mail: [keijo.lukkarinen@aanekoski.fi](mailto:keijo.lukkarinen@aanekoski.fi)

© Springer Nature Switzerland AG 2022  
T. Tuovinen et al. (eds.), *Computational Sciences and Artificial Intelligence in Industry, Intelligent Systems, Control and Automation: Science and Engineering 76*,  
[https://doi.org/10.1007/978-3-030-70787-3\\_13](https://doi.org/10.1007/978-3-030-70787-3_13)

## 1 Introduction

Obesity has become a global problem that concerns most countries nowadays. This development has led to severe individual and societal consequences in terms of obesity-related morbidity and mortality (GBD [5]). Several modifiable factors, such as eating habits and physical activity, explaining overweight and obesity are well understood. It is, however, still difficult to predict the individual's weight development in early childhood. The early identification of the individuals at the risk of developing obesity in later life could enable more effective prevention and interventions in primary health care. A reliable prediction model that may eventually be developed could be a valuable decision support tool for healthcare professionals.

The internationally accepted threshold values for adult overweight and obesity are defined by body mass index (BMI) cut-points of 25 kg/m<sup>2</sup> and 30 kg/m<sup>2</sup>, respectively (Cole et al. [2]). The BMI is calculated using the standard formula of weight (kg) divided by the square of body height (m). However, these fixed cut-point values are not directly applicable to children as the values are both lower and increase with age during maturation due to changes in body composition (Shields [15]). In order to use these static overweight and obesity threshold values for children, growth references that include BMI-for-age, such as de Onis et al. [3], should be employed.

Several previous studies indicate that machine learning-based predictive modeling approach, in which models are validated on independent data not included in the training set, may have potential in the task of overweight/obesity prediction (Morandi et al. [9]; Santorelli et al. [14]; Weng et al. [17]; Redsell et al. [12]; Graversen et al. [6]; Zhang et al. [19]; Dugan et al. [4]). Most of the existing studies have applied the logistic regression to the task, but only a few have applied machine learning methods of greater learning capacity (e.g., Zhang et al. [19]; Dugan et al. [4]). The model performance is typically reported in terms of cross-validated Area Under the Receiver Operating Characteristics Curve (AUC) and specificity and sensitivity measures. In contrast to predictive modeling, explanatory modeling has been employed extensively in the context of childhood obesity research (Simmonds et al. [16]). However, explanatory modeling is generally more focused on explaining the phenomenon as opposed to directly predicting the outcome, and explanatory approaches do not validate their results using independent data (Breiman [1]).

Several logistic regression models have demonstrated a fair (Morandi et al. [9]; Weng et al. [17]; Redsell et al. [12]) or good (Santorelli et al. [14]) level of AUC performance. AUC of 0.886 was reported for a model in which the obesity status at the age of two years was predicted from the birth weight and the weight gain during the first 12 months (Santorelli et al. [14]). The model was trained and validated with 731 and 867 cases, respectively.

Only a few overweight/obesity prediction results are available for more complex non-linear machine learning models (Zhang et al. [19]; Dugan et al. [4]). Zhang et al. [19] reported higher values for specificity (96% for a Naïve Bayes model) when compared with logistic regression models, but the sensitivity of the models was modest (60% for a Support Vector Machine model). Dugan et al. [4] evaluated

( $N = 7519$ ) several machine learning models using a significantly greater number of predictor variables (in total 167) than Zhang et al. [19]. The predictor variables were collected before the children's second birthday, and the task was to predict whether a child was obese at any point in time in the age ranging from two to ten years. The ten-fold cross-validation analysis indicates that classification performance in the obesity prediction task can be improved significantly by including more predictor variables. Sensitivity and specificity values for an ID3 decision tree model were 89% and 83%, and for a random tree 88% and 80%, respectively. Hence, the performance of these models clearly outperforms the ones presented by Zhang et al. [19]. Neither of these studies reported the performance in terms of AUC values. However, one must keep in mind that the results of these two studies are not directly comparable since the target range in Dugan et al. [4] study is significantly broader.

The studies above provide clear evidence of the potential of the data-driven predictive modeling approach in the overweight/obesity prediction. It is, however, important to bear in mind that, in addition to the chosen machine learning method, the quality of prediction results is dependent on several other factors, such as the chosen predictor variables, quality and amount of data, hyperparameter optimization, and the model assessment and selection. Therefore, it is not possible to draw conclusions about the mutual superiority of the proposed methods. However, it is essential to investigate how the performance of the proposed approaches and models can be generalized to new data sets.

The aim of this study is to investigate the potential of Finnish childhood growth data in building machine learning models for predicting overweight and obesity in later life. To achieve our goal, we gathered a set of existing studies in Rautiainen and Äyrämö [11] providing appropriate information for reproducing and cross-validating the models on Finnish data.

The remaining part of the study proceeds as follows: Sect. 2 is concerned with the materials and methods employed during the study, while Sect. 3 presents the prediction results. The study is discussed and concluded in Sects. 4 and 5.

## 2 Materials and Methods

### 2.1 Data

The original data set consisted of body weight and height values measured from 14197 people by health care specialists in the town of Äänekoski between the years 1986 and 2018. The dates of birth and measurement were also included. The original data included measurements from people of all ages that visited health care during that time, including those born before 1986. This data is referred to as “*Finnish data*” from here on.

Before training the models, all obviously incorrect (e.g., the date of measurement in the future), inconsistent (e.g., the date of measurement before the date of birth),



**Table 1** Previous predictive modeling settings selected for the evaluation in the present study (Rautiainen and Äyrämö [11])

Setting	Study	Description of setting
zhang_6w	Zhang et al. [19]	Predicting overweight (incl. obesity) on three-year-old children based on data recorded up to six weeks of age.
zhang_8m	Zhang et al. [19]	Predicting overweight (incl. obesity) on three-year-old children based on data recorded up to eight months of age.
zhang_2y	Zhang et al. [19]	Predicting overweight (incl. obesity) on three-year-old children based on data recorded up to two years of age.
dugan_u2y	Dugan et al. [4]	Predicting obesity on 2–10-year-old children based on data recorded up to under two years of age. If the child is obese at any age between 2–10 years, he/she is classified as obese for the purpose of the analysis.
weng_1y	Weng et al. [17]	Predicting overweight (incl. obesity) on three-year-old children based on data recorded up to one year of age.
redsell_1y	Redsell et al. [12]	Predicting overweight (incl. obesity) on five-year-old children based on data recorded up to one year of age.
graversen_b	Graversen et al. [6]	Predicting overweight (incl. obesity) on 13–16-year-old children based on data recorded at the birth of the child.
graversen_5y	Graversen et al. [6]	Predicting overweight (incl. obesity) on 13–16-year-old children based on data recorded up to five years of age.
graversen_8y	Graversen et al. [6]	Predicting overweight (incl. obesity) on 13–16-year-old children based on data recorded up to eight years of age.

or incomplete (e.g., either bodyweight or height or both are missing) values and duplicate data were removed from the data set. All the body weight and height values were then converted to kilograms and meters, respectively.

BMI values less than  $5 \text{ kg/m}^2$  and greater than  $55 \text{ kg/m}^2$  were considered invalid, and corresponding measurements were removed from the data. Moreover, at least six monotonically increasing values for the body height were required for each person to be included in the data set(s).

As a result of data cleaning, the final data set consisted of 6493 individual time series. The average length of the measurement intervals was 265 ( $\pm 173$  std) days. The average number of measurement points per person was 15.46 ( $\pm 6.84$  std). It is important to notice that the frequency of the measurements is considerably higher during the first years after the birth.

At this stage, all the time series were resampled to 1-d intervals with a linear interpolation method. After that, the final data sets were created by sampling the BMI time series at 30-d intervals starting from day number thirty and ending by the day defined in the research setting. Hence, the length of each time series included in the phase of model training varied between different research settings. Altogether nine research settings from five studies (Rautiainen and Äyrämö [11]) were chosen for evaluation on Finnish data (see Table 1). The research settings defined the input BMI time series and the age where the discrete target variables (overweight/obesity status) were determined. All except one research setting aimed to predict overweight (including obesity) status. The only exception was *dugan\_u2y*, which was aimed to predict obesity.

In total, there were nine data sets consisting of (1) BMI time series training data, (2) BMI time series testing data, and (3) response variable data, meaning that we had

one data set for each of the main research settings of the original studies examined. In addition, females and males were modeled separately in one study (Graversen et al. [6]), making the total number of models formed in this study 12. We already briefly discussed the input data used in the previous studies in Sect. 1. More detailed descriptions about the input data used in the original studies can be found in our earlier review Rautiainen and Äyrämö [11]. To summarize, none of the original studies employed time series data as input in the way this study does. Instead, they generally employed weight, height, or BMI information from a few specific time points.

For the response, we created a binary target variable with two variations (*not overweight vs. overweight* and *not obese vs. obese*) depending on the research setting. For all research settings except *dugan\_u2y*, the task was to predict the overweight (including obesity) status. The overweight and obesity thresholds were based on the age-adjusted ISO-BMI values defined for Finnish children aged between two and 18 years (Saari et al. [13]).

## 2.2 Model Fitting

Before fitting the models, the time series data was split into training (66.7%) and test sets in a ratio of 66.7/33.3. Stratification was employed in the sampling to keep the distribution of responses similar in the two data sets. The models were then fitted using several types of machine learning methods: logistic regression with  $l_1$ - and  $l_2$ -penalization (LogReg- $l_1$  and LogReg- $l_2$ , respectively),  $k$ -nearest-neighbor classifier (kNN) on principal components (PC), support vector machine (SVM) trained on the original time series and PCs, decision tree (DT), and neural networks (NN).

All the models were trained using a set of training samples  $\mathcal{D}_{\text{tr}} = \{\mathbf{x}_i, y_i\}_{i=1}^n$ , where  $\mathbf{x}_i \in \mathbb{R}^d$  is the vector of BMI values for  $i$ th children and  $y_i$  is a binary response variable created earlier. In addition,  $d$  is determined by the number of interpolated BMI values, and its value changes depending on the research setting employed. For example,  $d$  is equal to 12 when data up to one year is used because there are twelve sampling points (at 30, 60, 90, ..., 360 d since birth) in total.

*Logistic regression* is a simple linear classification method that is applied to predict the probabilities of the possible outcomes for given input data (Hastie et al. [7]). In a binary case, the model is defined as the posterior probability of the response variable

$$p(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(\mathbf{x}^T \boldsymbol{\beta} + \beta_0)},$$

where  $\boldsymbol{\beta} \in \mathbb{R}^d$  and  $\beta_0 \in \mathbb{R}$  are the model parameters determining the separating hyperplane between the classes in the input space. We applied  $l_1$ - and  $l_2$ -regularized logistic regression models that can be derived from the family of log-likelihood minimization problems (see, e.g., Hastie et al. [8]):



$$\min_{\beta_0, \beta} -l(\beta_0, \beta) + \frac{\lambda}{q} \|\beta\|_q^q,$$

where  $l$  is the binomial log-likelihood function. The hyperparameter  $\lambda$  is a positive constant determining the amount of regularization and must be optimized in a separate loop.  $l_1$ - and  $l_2$ -penalized logistic regression problems are obtained by choosing either  $q = 1$  or  $q = 2$ , respectively.

*Decision trees* are non-parametric classifiers in which the model is trained by inferring a set of simple decision rules from the predictor variables (Hastie et al. [7]). The basic idea is to fit a tree model that recursively partitions the input data space containing a set of data points into subregions using axis-parallel hyperplanes. The class label of a new data point can be predicted by recursively evaluating split decisions until the leaf node. Key issues to be determined before training a decision tree model are the criteria for splitting the training data into subsets at each node and for stopping the recursive splitting procedure (Zaki and Meira [18]). In this study, we applied an optimized version of the CART algorithm available at the Scikit-learn library. Gini index was chosen as the measure for the quality of a split, and it is defined by

$$G(D) = 1 - \sum_{k=1}^K p_k^2,$$

where  $p_k$  is the proportion of observations belonging to class  $k$  in a data set  $D$ . The following decision tree hyperparameters were optimized using the given options with the grid search:

- The number of predictor variables that were considered when determining the best split in an internal node:  $d$  and  $\sqrt{d}$  where  $d$  is the total number of predictor variables.
- The maximum depth of the tree was optimized in the set  $\{1, \dots, 15\}$ .
- The minimum percentage of samples required to split an internal node 10.0, 32.5, 55.0, 77.5, and 100.0%.
- The minimum percentage of samples required in a leaf node (a split point was not considered if there were fewer training samples than the given minimum value in any subsequent leaf nodes): 10.0, 20.0, 30.0, 40.0, and 50.0%.

*Support vector machines* are effective classification methods for high-dimensional problems (see, e.g., Hastie et al. [7]). SVMs aim for low generalization error by fitting a hyperplane to the data with the maximal distance to the nearest training samples of any class. Fitting an SVM model is a convex quadratic problem with linear inequality constraints defined in the primal form:

$$\begin{aligned} & \min_{\beta, \beta_0, \zeta_i} \frac{1}{2} \|\beta\|^2 + C \sum_{i=1}^n \zeta_i & (1) \\ \text{subject to} & \begin{cases} y_i(\beta^T \phi(\mathbf{x}_i) + \beta_0) \geq 1 - \zeta_i, \\ \zeta_i \geq 0, \quad i = \{1, \dots, n\}, \end{cases} \end{aligned}$$

where the model parameters  $\beta$ ,  $\beta_0$  define the hyperplane,  $C \geq 0$  is a regularization parameter for controlling overfitting vs. smoothness of the model in non-separable cases (the classes overlap in the input space),  $\zeta_i$ s is a slack variable giving the proportional amount by which  $i$ th prediction is on the wrong side of its margin, and  $\phi(\mathbf{x})$  is a basis function for producing a non-linear decision function.

The dual formulation of the cost (1) leads to a simpler convex problem (see Hastie et al. [7])

$$\begin{aligned} & \min_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j) \\ \text{subject to} & \begin{cases} \sum_{j=1}^n y_j \alpha_j = 0, \\ 0 \leq \alpha_i \leq C, \quad i = \{1, \dots, n\}, \end{cases} \end{aligned}$$

where the vector  $\alpha$  consists of the model coefficients determining the support vectors (see, e.g., Hastie et al. [7], for a more detailed derivation of the dual form), the kernel function  $K(\mathbf{x}_i, \mathbf{x}_j) = \phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$  computes the inner product in the enlarged space, and  $C$  is again the regularization parameter (the smaller the values of  $C$ , the smoother the class boundaries).

Both  $C$  and  $\phi(\mathbf{x})$  are considered hyperparameters, and they were optimized through grid search. The grid search space for SVM consisted of regularization parameter  $C = \{1, 2, \dots, 10\}$ , linear and radial basis functions, in addition to second, third, and fourth degree polynomial functions.

The decision function of an SVM model is

$$\hat{f}(\mathbf{x}) = \text{sgn} \left( \sum_{i=1}^n \hat{\alpha}_i y_i K(\mathbf{x}_i, \mathbf{x}) + \hat{\beta}_0 \right),$$

where  $\hat{\alpha}_i$  and  $\hat{\beta}_0$  are estimated on the training data.

In addition to the methods mentioned above, k-nearest-neighbor classifier (kNN) on principal components, neural networks, and gradient boosting classifiers were trained, and their performance was evaluated. However, since they did not produce the highest scores on any of the settings, we skip their more detailed description.

### 2.3 Performance Evaluation

For each machine learning method, the hyperparameters were optimized on the training set using the grid search strategy and stratified ten-fold cross-validation with F1-score as a measure of classifier performance. With all the fixed hyperparameters, the default values determined by the Scikit-learn library were used.

F1-score is defined by

$$F1 = 2 * \frac{\text{precision} * \text{sensitivity}}{\text{precision} + \text{sensitivity}}, \quad (2)$$

where

$$\text{precision} = \frac{TP}{TP + FP},$$

and

$$\text{sensitivity} = \frac{TP}{TP + FN}. \quad (3)$$

Here, TP (true positives) is the number of correctly detected true *overweight* or *obese* cases, FP (false positives) is the number of incorrectly detected *overweight* or *obese* cases, and FN (false negatives) is the number of incorrectly detected *not overweight* or *not obese* cases. F1-score is designed to balance the precision and sensitivity of the classifier by computing their harmonic mean. Precision is therein defined as the ratio of correct true positive (e.g., *overweight/obese*) predictions to the total number of positive cases in data. Finally, sensitivity is defined as the fraction of correct predictions with respect to all points in the positive class (Zaki and Meira [18]).

After fixing the optimal hyperparameters for each method, the final models were fitted on the training sets. The predictive accuracy of each model was then evaluated on the hold-out test sets. The results are reported in terms of sensitivity (3), F1-score (2), and specificity that is given by

$$\text{specificity} = \frac{TN}{TN + FP},$$

where TN (true negatives) is the number of correctly detected true *not overweight/not obese* cases and FP as given above. Similar to sensitivity, specificity is defined as the fraction of correct predictions with respect to all points in the negative (e.g., *not overweight/not obese*) class (Zaki and Meira [18]).

Next, the results are presented as the average values with standard deviations.

## 3 Results

All the models were trained and tested using functions available at Scikit-learn library version 0.20.1 (Pedregosa et al. [10]). The different research settings from the origi-

**Table 2** Predictive performance measures reported by the previous studies (Rautiainen and Äyrämö [11])

Setting	Method	N (Train)	N (Test)	Sensitivity (%)	Specificity (%)
zhang_6w	Naïve Bayes	11070	≈5400–5600	11.2	96.0
zhang_8m	Naïve Bayes	11070	≈5400–5600	35.5	91.5
zhang_2y	Naïve Bayes	11070	≈5400–5600	54.7	93.1
dugan_u2y	ID3 DT	6767	752	89.0	83.0
weng_1y	LogReg	8299	1715	76.9	66.5
redsell_1y	LogReg	8299	980	53.0	71.0
graversen_b	LogReg	4111	5414	24.0 (females) 17.4 (males)	92.1 (females) 91.7 (males)
graversen_5y	LogReg	4111	5414	38.9 (females) 28.2 (males)	94.4 (females) 94.2 (males)
graversen_8y	LogReg	4111	5414	49.2 (females) 38.7 (males)	96.0 (females) 96.7 (males)

**Table 3** Predictive performance measures obtained using research settings from the previous studies and Finnish data set

Setting	Method	N (Train)	N (Test)	Sensitivity (%)	Specificity (%)	F1-score
zhang_6w	LogReg-l2	1828	913	70.0	58.8	0.45
zhang_8m	DT	1828	913	63.7	77.6	0.53
zhang_2y	DT	1828	913	74.5	88.7	0.70
dugan_u2y	DT	2241	1120	45.8	95.1	0.49
weng_1y	LogReg-l2	1828	913	75.0	79.4	0.61
redsell_1y	LogReg-l2	1703	851	68.7	74.0	0.52
graversen_b*	LogReg-l2 (females)	485 (f)	243 (f)	55.6 (f)	60.0 (f)	0.41 (f)
	SVM (males)	522 (m)	261 (m)	97.9 (m)	1.2 (m)	0.52 (m)
graversen_5y	LogReg-l2	485 (f)	243 (f)	65.1 (f)	84.4 (f)	0.62 (f)
		522 (m)	261 (m)	62.8 (m)	81.4 (m)	0.64 (m)
graversen_8y	LogReg-l2 (females)	485 (f)	243 (f)	76.2 (f)	83.9 (f)	0.69 (f)
	SVM (males)	522 (m)	261 (m)	75.5 (m)	82.0 (m)	0.73 (m)

\* For the *graversen\_b* setting, the input data was the shortest available time series ( $d = 1$ ), at 30d of age

nal studies, as mentioned earlier, are presented in Table 1. Furthermore, the available measures from the original studies are shown in Table 2, and the performance measures for the best-obtained models on Finnish data are presented in Table 3. The parameters for methods found during the grid search are presented in Table 4.

The overall prediction performance of the models trained and tested on Finnish data was measured by F1-score. The F1-scores ranged from moderate (0.41 for the female LogReg-l2 model with *graversen\_b* setting) to good (0.73 for the male SVM model with *graversen\_8y* setting). The average F1-score was  $0.58 \pm 0.10$ . An explicit comparison of the F1-scores with the original studies is not possible because they were not reported, but we compared sensitivity and specificity values. The sensitivity of the Finnish models ranged from 45.8 to 97.9%, with an average of

**Table 4** Parameters for the employed methods found using grid search

Setting	Method	Parameters
zhang_6w	LogReg-l2	C (inverse of regularization strength): 0.25, solver: Stochastic average gradient descent (sag)
zhang_8m	DT	max_depth (the maximum depth of the tree): 1, max_features (the number of features to consider when looking for the best split): None, min_samples_leaf (the minimum number of samples required to be at a leaf node): 0.3, min_samples_split (the minimum number of samples required to split an internal node): 0.1
zhang_2y	DT	max_depth: 6, max_features: auto, min_samples_leaf: 0.2, min_samples_split: 0.1
dugan_u2y	DT	max_depth: 1, max_features: None, min_samples_leaf: 0.1, min_samples_split: 0.1
weng_1y	LogReg-l2	C: 1.5, solver: Newton method (newton-cg)
redsell_1y	LogReg-l2	C: 0.5, solver: Newton method (newton-cg)
graversen_b (females)	LogReg-l2	C: 0.25, solver: Stochastic average gradient descent (sag)
graversen_b (males)	SVM	C: 1, kernel: poly, degree (degree of the polynomial kernel function): 4
graversen_5y (females)	LogReg-l2	C: 1.5, solver: Newton method (newton-cg)
graversen_5y (males)	LogReg-l2	C: 0.5, solver: Stochastic average gradient descent (sag)
graversen_8y (females)	LogReg-l2	C: 1.5, solver: Newton method (newton-cg)
graversen_8y (males)	SVM	C: 2, kernel: linear

69.2 ± 12.2%. In the original studies, the average of sensitivity measures is clearly lower at 43.1 ± 22.1%.

Overall, in 11 out of 12 settings in the models using Finnish data, the estimated sensitivity was higher than in the original studies. Also, in one of the remaining cases, the *weng\_1y* setting, the sensitivity of the original model (76.9%) was nearly equal to our result (75.0%). Moreover, in this case, the specificity (79.4%) of the Finnish model is higher in comparison with the original study (66.5%) indicating better overall prediction performance on the Finnish data. The average specificity value for the Finnish models was 72.2 ± 23.7%, with a large range of values from 1.2% to 95.1%, whereas in the original studies, the average is higher 88.9 ± 9.7% and the range 66.5–96.7%.

For one setting, *redsell\_1y*, where the overweight (including obesity) status at the age of five was predicted from values measured during the first 12 months from the birth, the Finnish LogReg-l2 model produced both higher sensitivity (68.7% versus 53.0%) and specificity (74.0 and 71.0%) than reported for the original LogReg model. In all the other cases, greater sensitivity is obtained with the cost of lower specificity or vice versa.

## 4 Discussion

The overall prediction performance, measured by F1-score, of the models trained and tested on Finnish data varied from moderate to good in the range between 0.41 and 0.73 (the SVM model for males with *graversen\_8y* setting). In the absence of estimates for overall prediction performance, the measures cannot be directly compared with the previous models. However, the results show that six out of 12 F1-scores exceeded the level of 0.6, suggesting that also Finnish data may contain strong dependencies that can be utilized for building overweight and obesity prediction models.

The best overall prediction performance (F1-score = 0.73) was obtained by the male SVM model and the *graversen\_8y* setting. In this setting, the predictor time series covers the longest period from birth to the age of eight. The long history data likely explains a large part of the good overall performance compared with the other settings on Finnish data. The same age period with the Finnish female data also produced a relatively high F1-score (0.69) that is the third highest of all the cases.

The poorest performance was obtained by the LogReg-l2 model trained for females employing the *graversen\_b* setting. This poor performance can also be considered as an expected result since the *graversen\_b* setting is aimed at predicting the overweight/obesity at the age between 13 and 16 years from the measurements conducted at the birth of the child. Therefore it is the most extended prediction range of all the selected settings in this study.

By summarizing the original studies (Table 2) as a whole, it is not possible to observe such a clear effect by the length of the predictor time series. In the settings investigated by Graversen et al. [6], one can, however, observe an increasing trend in both sensitivity and specificity measures with the length of the predictor time series. The observations from the present results and the study by Graversen et al. [6] underline the fact that the early and accurate detection of the risk of childhood obesity are obviously conflicting objectives whose optimization requires most likely more versatile information from the early years of childhood. In all the settings evaluated in Graversen et al. [6], the detection rate is moderate or even poor, which means that the overall prediction performance of the models may not be sufficient.

Overall, the predictive performance of the models trained on Finnish data is well comparable with the performance measures reported in the original studies. Sensitivity in the Finnish tests was on average and in all except one case higher than the original studies.

The Finnish data outperformed the test results reported for the original model in one setting, *redsell\_1y*, both in sensitivity and specificity, indicating that there is information in Finnish data. Since the original study Redsell et al. [12] employed also other variables in addition to weight information, such as maternal and paternal information, the better result achieved here might suggest that utilizing more precise time series data for carefully chosen variable(s) could be advantageous.

## 5 Conclusion

This study has shown that Finnish data can be utilized in the task of overweight/obesity prediction. One potential way of further enhancing the performance of the models might be to employ additional predictors to existing data. Further research could be conducted to determine the potential of models with greater learning capacity, such as recurrent neural networks. It might also be useful to consider further model building practices, i.e., predictor selection and significance of prediction (Rautiainen and Äyrämö [11]).

**Acknowledgements** Ilkka Rautiainen received funding from Business Finland and a grant from the Jenny and Antti Wihuri Fund.

## References

1. Breiman L (2001) Statistical modeling: the two cultures. *Stat Sci* 16(3):199–215. <https://doi.org/10.1214/ss/1009213726>
2. Cole TJ, Bellizzi MC, Flegal KM, Dietz WH (2000) Establishing a standard definition for child overweight and obesity worldwide: international survey. *BMJ* 320(7244):1240. <https://doi.org/10.1136/bmj.320.7244.1240>
3. de Onis M, Onyango AW, Borghi E, Siyam A, Nishida C, Siekmann J (2007) Development of a WHO growth reference for school-aged children and adolescents. *Bull World Health Organ* 85(9):660–667. <https://doi.org/10.1590/S0042-96862007000900010>
4. Dugan TM, Mukhopadhyay S, Carroll A, Downs S (2015) Machine learning techniques for prediction of early childhood obesity. *Appl Clin Inform* 6(3):506–520. <https://doi.org/10.4338/ACI-2015-03-RA-0036>
5. GBD (2017) Health effects of overweight and obesity in 195 countries over 25 years. *New Engl J Med* 377(1):13–27. <https://doi.org/10.1056/NEJMoa1614362>. (Written by GBD 2015 Obesity Collaborators)
6. Graversen L, Sørensen TIA, Gerds TA, Petersen L, Sovio U, Kaakinen M, Sandbaek A, Laitinen J, Taanila A, Pouta A, Järvelin M-R, Obel C (2015) Prediction of adolescent and adult adiposity outcomes from early life anthropometrics. *Obesity* 23(1):162–169. <https://doi.org/10.1002/oby.20921>
7. Hastie T, Tibshirani R, Friedman J (2009) *The elements of statistical learning: data mining, inference and prediction*, 2nd edn. Springer, Berlin. <http://www-stat.stanford.edu/~tibs/ElemStatLearn/>
8. Hastie T, Tibshirani R, Wainwright M (2015) *Statistical learning with sparsity: the lasso and generalizations*. Chapman and Hall/CRC, Boca Raton
9. Morandi A, Meyre D, Lobbens S, Kleinman K, Kaakinen M, Rifas-Shiman SL, Vatin V, Gaget S, Pouta A, Hartikainen A-L, Laitinen J, Ruukonen A, Das S, Khan AA, Elliott P, Maffei C, Gillman MW, Järvelin M-R, Froguel P (2012) Estimation of newborn risk for child or adolescent obesity: lessons from longitudinal birth cohorts. *PLoS One* 7(11):e49919. <https://doi.org/10.1371/journal.pone.0049919>
10. Pedregosa F, Varoquaux G, Gramfort A, Michel V, Thirion B, Grisel O, Blondel M, Prettenhofer P, Weiss R, Dubourg V, Vanderplas J, Passos A, Cournapeau D, Brucher M, Perrot M, Duchesnay E (2011) Scikit-learn: machine learning in Python. *J Mach Learn Res* 12:2825–2830
11. Rautiainen I, Äyrämö S (2019) Predicting overweight and obesity in later life from childhood data: a review of predictive modeling approaches. [arXiv:1911.08361](https://arxiv.org/abs/1911.08361)



12. Redsell SA, Weng S, Swift JA, Nathan D, Glazebrook C (2016) Validation, optimal threshold determination, and clinical utility of the infant risk of overweight checklist for early prevention of child overweight. *Child Obes* 12(3):202–209. <https://doi.org/10.1089/chi.2015.0246>
13. Saari A, Sankilampi U, Hannila M-L, Kiviniemi V, Kesseli K, Dunkel L (2011) New Finnish growth references for children and adolescents aged 0 to 20 years: length/height-for-age, weight-for-length/height, and body mass index-for-age. *Ann Med* 43(3):235–248. <https://doi.org/10.3109/07853890.2010.515603>
14. Santorelli G, Petherick ES, Wright J, Wilson B, Samiei H, Cameron N, Johnson W (2013) Developing prediction equations and a mobile phone application to identify infants at risk of obesity. *PLoS One* 8(8):e71183. <https://doi.org/10.1371/journal.pone.0071183>
15. Shields M (2006) Overweight and obesity among children and youth. *Health Rep* 17(3):27–42
16. Simmonds M, Burch J, Llewellyn A, Griffiths C, Yang H, Owen C, Duffy S, Woolacott N (2015) The use of measures of obesity in childhood for predicting obesity and the development of obesity-related diseases in adulthood: a systematic review and meta-analysis. *Health Tech Assess* 19(43):1–336. <https://doi.org/10.3310/hta19430>
17. Weng SF, Redsell SA, Nathan D, Swift JA, Yang M, Glazebrook C (2013) Estimating overweight risk in childhood from predictors during infancy. *Pediatrics* 132(2):e414–e421. <https://doi.org/10.1542/peds.2012-3858>
18. Zaki MJ, Meira W Jr (2014) *Data mining and analysis: fundamental concepts and algorithms*. Cambridge University Press, New York
19. Zhang S, Tjortjis C, Zeng X, Qiao H, Buchan I, Keane J (2009) Comparing data mining methods with logistic regression in childhood obesity prediction. *Inf Syst Front* 11(4):449–460. <https://doi.org/10.1007/s10796-009-9157-0>





### III

## **PRECISION EXERCISE MEDICINE: PREDICTING UNFAVOURABLE STATUS AND DEVELOPMENT IN THE 20-M SHUTTLE RUN TEST PERFORMANCE IN ADOLESCENCE WITH MACHINE LEARNING**

by


Ilkka Rautiainen, Laura Joensuu, Sami Äyrämö, Heidi J Syväoja, Jukka-Pekka Kauppi, Urho M Kujala, and Tuija H Tammelin 2021

BMJ Open Sport & Exercise Medicine, 7:e001053

<https://doi.org/10.1136/bmjsem-2021-001053>

Reproduced with kind permission of BMJ Publishing Group Ltd.

# Precision exercise medicine: predicting unfavourable status and development in the 20-m shuttle run test performance in adolescence with machine learning

Laura Joensuu <sup>1,2</sup>, Ilkka Rautiainen,<sup>3</sup> Sami Äyrämö,<sup>3</sup> Heidi J Syväoja,<sup>2</sup> Jukka-Pekka Kauppi,<sup>3</sup> Urho M Kujala,<sup>1</sup> Tuija H Tammelin<sup>2</sup>

**To cite:** Joensuu L, Rautiainen I, Äyrämö S, *et al.* Precision exercise medicine: predicting unfavourable status and development in the 20-m shuttle run test performance in adolescence with machine learning. *BMJ Open Sport & Exercise Medicine* 2021;**7**:e001053. doi:10.1136/bmjsem-2021-001053

► Additional supplemental material is published online only. To view, please visit the journal online (<http://dx.doi.org/10.1136/bmjsem-2021-001053>).

LJ and IR contributed equally.

Accepted 7 May 2021



© Author(s) (or their employer(s)) 2021. Re-use permitted under CC BY-NC. No commercial re-use. See rights and permissions. Published by BMJ.

<sup>1</sup>Faculty of Sport and Health Sciences, University of Jyväskylä, Jyväskylä, Finland

<sup>2</sup>LIKES Research Centre for Physical Activity and Health, Jyväskylä, Finland

<sup>3</sup>Faculty of Information Technology, University of Jyväskylä, Jyväskylä, Finland

**Correspondence to**  
Mrs Laura Joensuu;  
[laura.p.joensuu@jyu.fi](mailto:laura.p.joensuu@jyu.fi)

## ABSTRACT

**Objectives** To assess the ability to predict individual unfavourable future status and development in the 20m shuttle run test (20MSRT) during adolescence with machine learning (random forest (RF) classifier).

**Methods** Data from a 2-year observational study (2013–2015, 12.4±1.3 years, n=633, 50% girls), with 48 baseline characteristics (questionnaires (demographics, physical, psychological, social and lifestyle factors), objective measurements (anthropometrics, fitness characteristics, physical activity, body composition and academic scores)) were used to predict: (Task 1) unfavourable future 20MSRT status (identification of individuals in the lowest 20MSRT tertile after 2 years), and (Task 2) unfavourable 20MSRT development (identification of individuals with 20MSRT development in the lowest tertile among adolescents with baseline 20MSRT below median level).

**Results** Prediction performance for future 20MSRT status (Task 1) was (area under the receiver operating characteristic curve, AUC) 83% and 76%, sensitivity 80% and 60%, and specificity 78% and 79% in girls and boys, respectively. Twenty variables showed predictive power in boys, 14 in girls, including fitness characteristics, physical activity, academic scores, adiposity, life enjoyment, parental support, social status in school and perceived fitness.

Prediction performance for future development (Task 2) was lower and differed statistically from random level only in girls (AUC 68% and 40% in girls and boys).

**Conclusion** RF classifier predicted future unfavourable status in 20MSRT and identified potential individuals for interventions based on a holistic profile (14–20 baseline characteristics). The MATLAB script and functions employing the RF classifier of this study are available for future precision exercise medicine research.

## INTRODUCTION

Precision medicine is prevention and treatment strategies of diseases taking the individual variability into account.<sup>1</sup> Recently, a similar concept called precision exercise medicine was brought forward where the role of physical activity (PA) and cardiorespiratory

## Key messages

### What is already known

- The 20-m shuttle run test is commonly used in adolescents to estimate unfavourable cardiorespiratory fitness
- Currently used methods for assigning interventions based on the 20-m shuttle run test have limitations in individual level accuracy

### What are the new findings

- Machine learning algorithm was able to identify adolescents with unfavourable future 20 m shuttle run test (20MSRT) status based on 14 baseline characteristics in girls, and 20 in boys.
- This study provides an example with attached MATLAB script and functions how to use machine learning in precision exercise medicine.
- Adolescents' overall physical, psychological and social status are recommended to be assessed before deciding on interventions based on the 20MSRT score.

fitness (CRF) in health enhancement was acknowledged.<sup>2</sup> However, currently, the focus in precision exercise medicine is mainly on exploring treatment procedures and exercise response variability in adults.<sup>2–3</sup> Nevertheless, many chronic diseases have origins already in early childhood.<sup>4</sup> Prevention strategies warrant more focus on children and adolescents, especially as health risks have associations with CRF<sup>5</sup> and reversibility with exercise interventions in this age group.<sup>6</sup>

The 20-m shuttle run test (20MSRT) is the most commonly used field test to estimate CRF.<sup>7</sup> Low 20MSRT score has adverse associations with many aspects of children's and adolescents' daily lives. Previous studies have reported 20MSRT associated with lower overall physical performance,<sup>8</sup> poorer tissue health (including adiposity,<sup>8</sup> brain<sup>9</sup> and bone tissue<sup>10</sup>), lower cardiometabolic and psychosocial health, and cognitive performance.<sup>8</sup>



However, currently used methods to assign interventions based on the 20MSRT have limitations by their individual level accuracy.<sup>7,11</sup> The ability to predict 20MSRT prospects during adolescence would enhance the identification of potential individuals for lifestyle interventions.

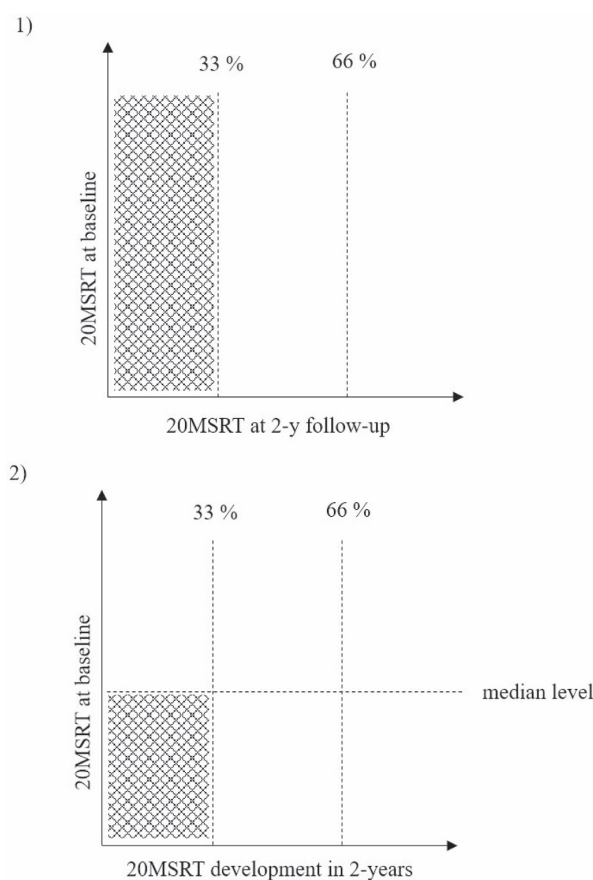
Machine learning (ML)-based pattern recognition approaches have emerged as promising alternatives to traditional statistical methods in precision exercise medicine.<sup>3</sup> Random forest (RF) is a commonly used ML algorithm. Contrary to other high learning capacity methods, such as neural networks and support vector machines, major advantages of RF include that the extensive tuning of hyperparameters is not required and overfitting the model is usually of lesser concern. An additional benefit especially suited for our research goals is extracting the estimates of importance for each variable in the data.<sup>12,13</sup> The main aim of this study was to evaluate the performance of RF on predicting future individual unfavourable 20MSRT status and development during adolescence based on 48 baseline variables, including physical, psychological and social indicators. Two prediction tasks were implemented: (Task 1) prediction of unfavourable future 20MSRT status (identification of individuals in the lowest 20MSRT tertile after 2 years), and (Task 2) prediction of unfavourable 20MSRT development in adolescents with limitations in their 20MSRT performance (identification of individuals with 20MSRT development in the lowest tertile among adolescents with baseline 20MSRT below median level). Task 1 focuses on the normal population, while Task 2 focuses specifically on children and adolescents who are more likely to experience the adverse outcomes related to lower 20MSRT performance.

We hypothesised that the baseline data contain variables that can predict future 20MSRT status and development. A secondary aim was to evaluate with a data-driven approach the best predictors of unfavourable 20MSRT prospects out of a wide range of baseline characteristics. We furthermore provide the predictive modelling algorithms used in this study for future research.

## METHODS

### Study design and participants

Secondary data analyses were performed for data collected in a 2-year longitudinal observational study (2013–2015) related to the Finnish Schools on the Move programme.<sup>14</sup> Data contained information from 971 students (mean 12.5±1.3 years, min 9.2 years, max 15.3 years, 52% girls). The sample of this study was further reduced to 633 (50% girls) (Task 1) and 300 subjects (50% girls) (Task 2), described in more detail in the Predictive modelling section. The data were collected at baseline during Spring and Fall semesters (1 May 2013 and 8 November 2013) and at follow-up during the Spring semester (1 May 2015) in nine Finnish public schools. The baseline and follow-up measurements during the Spring semester were performed within the same calendar week in each school.



**Figure 1** Prediction tasks were (A) unfavourable future 20MSRT status (identification of individuals in the lowest 20MSRT tertile after 2 years), and (B) unfavourable 20MSRT development in adolescents with limitations in their 20MSRT performance (identification of individuals with 20MSRT development in the lowest tertile among adolescents with baseline 20MSRT below median level). Both of these target tertile groups are highlighted in grey. The exact outcome variables to be predicted were (A) status of 20MSRT at follow-up (laps) and (B) absolute change between baseline and follow-up (in laps). The median level refers to the 50% performance level that was determined for each age cohort and both sexes separately to select the study sample in Task 2. The 33%, 66% cut-offs represent the tertiles used in Tasks 1 and 2. In both tasks, the outcome tertiles were determined for each age cohort and both sexes separately. 20MSRT, 20-m shuttle run test.

Forty-eight baseline variables (see the full list in online supplemental information document 1) were used in the prediction tasks (figure 1). Information regarding participants' demographics, physical, psychological and social factors was obtained from self-assessment questionnaires and non-invasive objective measurements.

### Self-assessment questionnaires

Participants completed two web-based questionnaires at baseline. Due to the extensiveness of the questionnaires, the data were collected in two parts: a first round

during the Spring 2013 and a second during the Fall 2013 semester (see division in online supplemental information document 1). In addition to basic demographic information (age and sex), the questionnaires assessed student's perceptions of their physical, psychological, and social status and health-related behaviour, for example, subjective evaluation of PA,<sup>15</sup> pubertal status on Tanner scale,<sup>16</sup> societal status of the family,<sup>17</sup> perceived health,<sup>18</sup> and cigarette, alcohol, and unhealthy food consumption.

### Objective measurements

All objective measurements were performed during the Spring semester of 2013. Body height was measured with an accuracy of 0.1 cm (Charder HM 200P scale). Body composition and mass were measured in light clothing using a bioelectrical impedance analysis device (InBody 720, Biospace Co.). Waist circumference was measured according to WHO guidelines.<sup>19</sup>

Physical fitness measurements were conducted in schools during the school day, with measurements included in the Finnish national Move!—monitoring system for physical functional capacity<sup>20</sup>: 20MSRT, push-up, curl-up, 5-leaps test, throwing–catching combination test and flexibility. Procedures for fitness measurements are described in detail in our previous baseline article.<sup>21</sup> The 20MSRT followed the Eurofit protocol and was recorded as laps run until voluntary exhaustion.

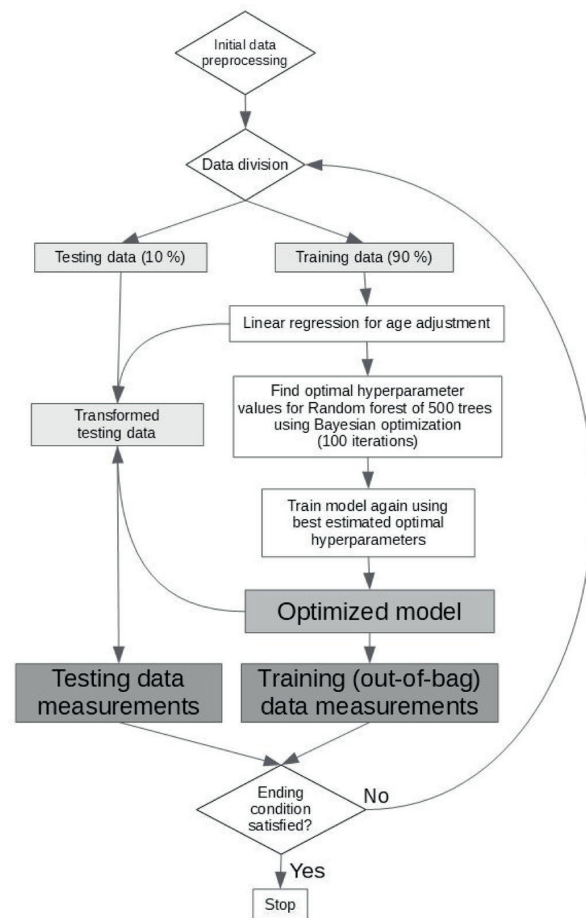
Device-based PA was evaluated using a hip-worn accelerometer (ActiGraph GT3X+, wGT3X+, Pensacola, Florida, USA) during a 7-day measurement period with raw 30 Hz acceleration, standard filtering and 15 s epoch conversion. Evenson criteria were used to define sedentary (<100 counts/min (cpm)), light (101–2295 cpm), moderate-to-vigorous (2296–20 000 cpm) physical activity (MVPA).<sup>22</sup> The valid amount of data was set for at least 500 min/day (between 07:00 and 23:00),<sup>23</sup> including at least 2 weekdays and 1 weekend day. Activity intensities were converted into weighted mean values per day (eg,  $MVPA = ((\text{average MVPA min/day of weekdays} \times 5 + \text{average MVPA min/day of weekend days} \times 2) / 7)$ ).

Academic scores (teacher-rated grade points) included grade point average (GPA) and grade point in physical education. Regional education services provided the data.

### Predictive modelling

The predictive modelling algorithms are provided in a data file (online supplemental information document 2) and available for future studies. All analyses were performed using MATLAB R2018a with the Statistics and Machine Learning Toolbox and conducted separately for both sexes.

The flow chart of predictive modelling is presented in figure 2. Please see the full details of the analyses in the online supplemental information document 3.



**Figure 2** The flow chart of predictive modelling.

#### Initial data preprocessing

##### Target variable formatting

The target variables to be predicted were (1) status of 20MSRT at follow-up and (2) absolute change in 20MSRT test result (laps) between the baseline and the follow-up (figure 1). The tertile groups were determined for both sexes and each age cohort separately. From a total of 971 observations, the 20MSRT baseline level could be determined for 871 students. A total of 633 participants were included in the Task 1 analysis. Exclusion criteria included participants with no result from the 20MSRT follow-up test. Here the missing mechanism was assumed to be missing completely at random. Altogether 300 adolescents were included in the Task 2 analyses. These participants had a recorded result for both 20MSRT tests, and their baseline 20MSRT result was below the age-specific and sex-specific median level. Here participants with no results from either of the two 20MSRT tests were excluded from the analysis.

Variables heavily dependent on age (see online supplemental information document 3 for a list) were age-adjusted using linear regression. The age-adjustment was first performed for the training data, and the residual





information was thereafter used to age adjust the corresponding variables in the testing data.

### Data division

The 10-fold cross-validation (CV) was used for model assessment where the data set (eg, in Task 2:  $n=150$  boys,  $n=150$  girls) was divided into 10 subsamples ( $n=15$  participants per subsample) called folds. Nine folds were then used as the training data (90% of the whole data set, to fit the tree model and estimate the variable importance values) and one fold as the testing data (10% of the whole data set, to evaluate the prediction accuracy on an independent sample). The procedures of training and prediction were then performed for these folds in a rotating manner, where eventually, all the folds had been used for training and testing. These procedures provided in total a set of 10 data-driven prediction models. The average performance of these 10 prediction models is shown in the Results section.

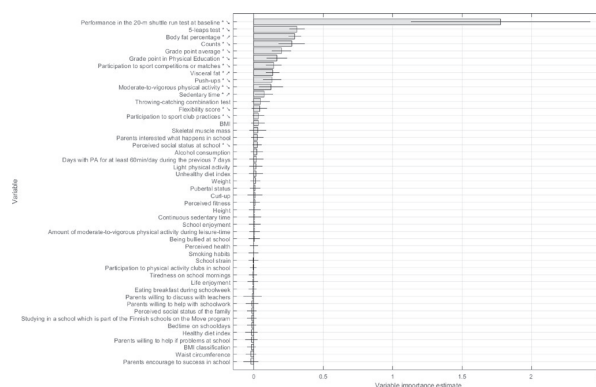
### Training and prediction

RF is an ML method that grows a forest of multiple de-correlated decision trees.<sup>13</sup> This forest of trees is thereafter employed as a voting ensemble, where each tree votes for the group of a single student (ie, does the individual belong to the lowest, middle or highest tertile group). The final predicted group for the student has the most votes in the whole forest.<sup>12, 13</sup> For each of the 10 folds, the trained model was employed to predict the testing portion of data. The area under the receiver operating characteristic curve (AUC), sensitivity and specificity metrics were recorded. A t-test in MATLAB was performed for AUC results to determine if the mean was significantly ( $p<0.05$ ) above the random level of 0.5.

The prediction strength of each feature is estimated using the out-of-bag (OOB) samples of each tree, that is, training data samples that have not been used when forming the tree. The OOB samples are shown to the tree, and the F1-score measure (online supplemental information document 3) of the predictions are recorded. Then the values of each feature are permuted one-by-one randomly, and after each permutation, the classification error is calculated again. This procedure is applied to all the trees in the forest. The final estimate of individual feature importance is the difference between the original classification error and the randomly permuted feature classification error, averaged for all the trees.<sup>12</sup> The final list of statistically significant ( $p<0.05$ ) predictors (online supplemental information document 5) was then formed, using MATLAB's t-test function. T-test was again performed for each predictor to determine which feature importance estimates were significantly above the mean of zero, indicating that they had predictive power.

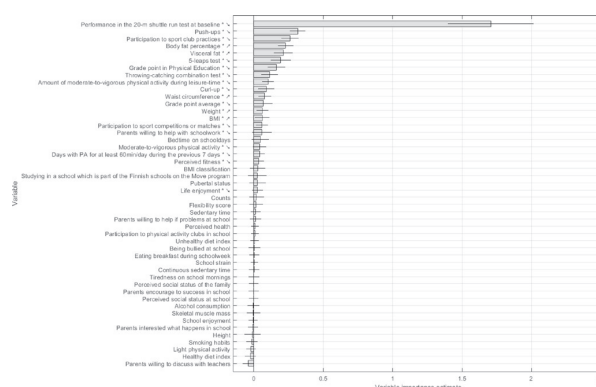
### The direction of the associations

The directions for the significant variables (significance set at  $p<0.05$ , presented in figures 3 and 4) were estimated using a separate receiver operating characteristic



**Figure 3** Best predictors for Task 1 in girls (20MSRT performance in the lowest tertile at 2-year follow-up). Statistically significant predictors are marked with \* ( $p<0.05$ ). Descending arrow ( $\searrow$ ): low values are associated with 20MSRT in the lowest tertile. Ascending arrow ( $\nearrow$ ): high values are associated with 20MSRT in the lowest tertile. The solid line represents the 95% CI. Variable importance estimate indicates the significance of the predictor. 20MSRT, 20-m shuttle run test.

(ROC) analysis.<sup>24</sup> The analysis was performed for the two prediction tasks, separately for girls and boys. Here, the whole data were employed without separation to training and testing data sets. Each variable in the data was then used one by one. The idea was to see how well a single variable can separate the data into two groups: the first group contained the lowest tertile and the second group contained the two upper tertiles. The separation threshold in the analysis is then changed step-by-step. At each step, two metrics needed for the ROC curve, sensitivity and specificity, are recorded. For each variable, we recorded the AUC value. The AUC value was then compared with



**Figure 4** Best predictors for Task 1 in boys (20MSRT performance in the lowest tertile at 2-year follow-up). Statistically significant predictors are marked with \* ( $p<0.05$ ). Descending arrow ( $\searrow$ ): low values are associated with 20MSRT in the lowest tertile. Ascending arrow ( $\nearrow$ ): high values are associated with 20MSRT in the lowest tertile. The solid line represents the 95% CI. Variable importance estimate indicates the significance of the predictor. 20MSRT, 20-m shuttle run test.

**Table 1** Descriptives of the study sample at baseline

	Boys (n=319)	Girls (n=314)
Age (years)	12.5±1.3	12.3±1.3
Height (cm)	156.1±11.7	154.1±9.6
Weight (kg)	46.1±12.9	44.8±10.5
BMI (kg/m <sup>2</sup> )	18.6±3.3	18.7±3.1
20MSRT (laps)	45.3±19.0	36.4±15.2
20MSRT centile*	60th	70th
MVPA (min/day)	58.0±22.4	48.3±17.9
Pubertal status <sup>†</sup>	2.6±1.0	2.5±0.9

Units are means and SD unless other mentioned.

\*International normative values by Tomkinson *et al*, 2016.

<sup>†</sup>Classification is based on self-assessment questionnaire and Tanner's scale.

BMI, body mass index; 20MSRT, 20-m shuttle run test; MVPA, accelerometry-based moderate-to-vigorous physical activity.

the random level (0.5). If the value was higher than the random level, we assumed that the variable information is applied correctly. The associated direction was that the higher the variable value, the higher the probability of the student belonging to the lowest tertile. Additionally, if the AUC value was lower than 0.5, a simple transformation of multiplying all the variable values with the number -1 was made, and the AUC was then calculated again. In this case, the associated direction was inverted: the lower the variable value, the higher the probability of belonging to the lowest tertile. The results of the ROC analysis are presented in online supplemental information document 4.

### Patient and public involvement

Patients or the public were not involved in designing, analysing or interpreting this study.

### RESULTS

The characteristics of the study sample are described in table 1. Participants' average performance in the 20MSRT was 45.3 and 36.4 laps at baseline in boys and girls, representing the 60th and 70th centile in the international normative values for 20MSRT.

### Prediction performance

The ability of the RF method to predict unfavourable future 20MSRT status (Task 1) is presented in table 2. The AUC values were higher in girls (0.83) than in boys (0.76), both statistically higher than the random level of 0.5 ( $p<0.001$ ). Sensitivity (individuals correctly predicted to belong to the lowest performance tertile) was higher in girls (0.80) than in boys (0.60). Specificity (individuals correctly predicted not to belong to the lowest performance tertile) was 0.78 in girls and 0.79 in boys.

The ability of the RF method to predict unfavourable 20MSRT development in a group of adolescents with baseline 20MSRT below the median level (Task 2) is presented in table 2. The prediction performance of ML was lower in these analyses. The AUC values were higher in girls (0.68) than boys (0.40), but only girls' predictions statistically differed from the random level ( $p=0.001$ ). Sensitivity (individuals correctly predicted to belong to the lowest development group) was higher in girls (0.59) than in boys (0.13). Specificity (individuals correctly predicted not to belong to the lowest development group) was 0.70 in girls and 0.79 in boys.

### Best predictors of 20MSRT prospects

The statistically significant predictors for Tasks 1 and 2 are represented in figures 3 and 4. The x-axis in the figures gives the estimate for variable importance, calculated using the increase or decrease in classification error when the predictor values are randomly permuted separately for each predictor. The higher the estimate, the higher is the significance of the predictor. Please see detailed information related to the direction and statistical significance of the variables in online supplemental information document 4. The top predictor for Task 1 was 20MSRT performance at baseline, both in boys and girls ( $p<0.001$ , figures 3 and 4), indicating that low initial 20MSRT performance predicts low performance also after 2 years.

Girls had 13 additional predictors (figure 3): low performance in other physical fitness tests (5-leaps test ( $p<0.001$ ), push-ups ( $p<0.001$ ) and flexibility score ( $p=0.049$ )), high markers of adiposity (body fat percentage ( $p<0.001$ ) and visceral fat ( $p<0.001$ )), low

**Table 2** The overall prediction performance of the unfavourable future 20MSRT status and development

	AUC	95% CI	P value	Sensitivity	95% CI	Specificity	95% CI
Task 1	Unfavourable future 20MSRT status (identification of individuals in the lowest 20MSRT tertile after 2 years)						
Girls	0.83	0.76 to 0.90	<0.001	0.80	0.69 to 0.91	0.78	0.74 to 0.82
Boys	0.76	0.71 to 0.81	<0.001	0.60	0.52 to 0.68	0.79	0.74 to 0.84
Task 2	Unfavourable 20MSRT development (identification of individuals with 20MSRT development in the lowest tertile among adolescents with baseline 20MSRT below median level)						
Girls	0.68	0.60 to 0.76	0.001	0.59	0.50 to 0.68	0.70	0.59 to 0.81
Boys	0.40	0.29 to 0.51	0.108	0.13	0.04 to 0.22	0.79	0.70 to 0.88

P value: statistical difference of the AUC value from the random level of 0.5; Sensitivity: individuals correctly predicted to belong to the explored group; Specificity: individuals correctly predicted not to belong to the explored group. AUC, area under the receiver operating characteristic curve; ;20MSRT, 20-m shuttle run test.



markers of PA (accelerometry-based counts ( $p < 0.001$ ), MVPA ( $p = 0.003$ ), participation to sport club practices ( $p = 0.025$ ) or competitions ( $p < 0.001$ ) and high percentage of accelerometry-based sedentary time ( $p = 0.009$ )), low academic scores (GPA and grade point in physical education (both  $p < 0.001$ )) and low perceived social status in school ( $p = 0.015$ ), all predicting placement in the lowest 20MSRT tertile after 2 years.

In addition to the baseline 20MSRT performance, boys had 19 additional predictors (figure 4): low performance in other physical fitness tests (push-ups ( $p < 0.001$ ), 5-leaps test ( $p < 0.001$ ), throwing-catching combination test ( $p < 0.001$ ) and curl-up ( $p = 0.001$ )), high markers of adiposity (body fat percentage ( $p < 0.001$ ), visceral fat ( $p < 0.001$ ), waist circumference ( $p < 0.001$ ), weight ( $p < 0.001$ ) and BMI ( $p = 0.005$ )), low academic scores (grade point in physical education ( $p < 0.001$ ), and GPA ( $p = 0.015$ )), low markers of PA (participation to sport club practices ( $p < 0.001$ ) or competitions ( $p = 0.001$ ), self-reported PA status (two questions:  $p < 0.001$  and  $p = 0.006$ ) and accelerometry-based MVPA ( $p = 0.020$ )), low parents' willingness to help with schoolwork ( $p = 0.045$ ), low perceived fitness ( $p = 0.007$ ) and low life enjoyment ( $p = 0.042$ ), all predicting future placement in the lowest 20MSRT performance tertile after 2 years.

As prediction performance for 20MSRT development was below 0.7 for both sexes, the best predictors are recommended to be interpreted with caution. These results are described in online supplemental information document 5.

## DISCUSSION

### Main findings

ML approach was able to predict, based on baseline characteristics, unfavourable future 20MSRT status with 0.76–0.83 (AUC) accuracy. Prediction performance was better in girls than in boys (eg, sensitivity values 0.80 in girls and 0.60 in boys). The prediction performance declined when predicting unfavourable 20MSRT development in a group of adolescents with an initial 20MSRT below the median level. These findings indicate that ML was able to identify potential individuals for interventions. Additionally, future fitness status might be easier to predict than development, at least in a group of adolescents with more homogeneous 20MSRT performance capacity.

### Best predictors of individual fitness development

Our findings showed that baseline 20MSRT performance was the best predictor of future performance in a large group of adolescents. However, this study highlighted 13–19 variables (out of 48 variables) with predictive power. These variables included a low performance in other field-based physical fitness tests, low perceived fitness, high markers of adiposity, low markers of PA, low academic achievement in school, low grade in physical education, low life enjoyment, low parental support and low perceived social status at school. These findings

indicate that multiple factors, that is, adolescents' overall physical, psychological and social well-being, contribute to the trajectory of the 20MSRT during adolescence. This information adds to the previous body of research where performance development is typically examined through growth and maturation ignited morphological changes.<sup>25</sup>

### Precision exercise medicine prospects

These promising findings also provide new prospects for precision exercise medicine in adolescents. Findings suggest that preventive measures linked to the 20MSRT score benefit from the ML-enabled holistic approach. In ML, patterns are explored from the data. This has benefits as data-driven characteristic profiles can be recognised if such exist in the data. Furthermore, the CV technique helps overcome a phenomenon where models or thresholds created with traditional statistics tend to fit poorly with other data sets or future individual observations.<sup>26</sup> An ML approach is recommended to be considered in future precision exercise medicine studies aiming to identify potential individuals for interventions.

Our findings indicated that information from adolescents' overall physical, psychological and social status provides additional value over evaluating only an individual's 20MSRT score. Potential use-cases are, for example, the national or regional fitness monitoring systems where a large number of children and adolescents are tested (up to >90% of age-cohort). Resources for interventions are typically limited and necessary to be directed for correct individuals. The next steps to use this method in practice would be to train the final model with selected feasible variables and to collect independent test data that the model could be evaluated against. To reduce the number of variables, for example, to indicate PA, it is possible to employ a stepwise variable elimination method to RF to select only the best variable.<sup>27</sup>

It is, however, important to use ML methods and computational power robustly. The availability of ML libraries and computational power lead easily to data fishing. This means that a fair application of CV techniques must assess the generalisation ability of the models, and the risk of chance findings should be eliminated using permutation testing or other relevant techniques. In the present framework, these aspects of ML application have been considered carefully.

### Strengths and limitations

The strengths of this study were the novel application for RF and the approach to predict individual fitness development in apparently healthy adolescents, the extensiveness of the variables in the data sample, robust analyses and measurements performed by educated professionals. Limitations include the 2-year duration of the study—more prominent changes could have potentially emerged with a longer follow-up period. The data sample was limited by its size (eg,  $n = 50$  in the lowest tertile in Task 2), possibly influencing prediction performance. There is also room for improvement in handling

the importance of variables. For example, it is possible to employ a stepwise variable elimination method to RF to reduce the effect of multicollinearity in data. The study used a sample from an observational study. Despite the efforts, sampling bias might exist and affect the generalisability of the findings to the adolescent population.

### Conclusion

With the ML approach, we could predict unfavourable future 20MSRT status based on 14–20 baseline characteristics and identify potential individuals for interventions. These promising findings support adopting a more holistic approach, taking physical and psychological and social factors into account in large-scale fitness monitoring systems. The ML algorithms used in this study are provided for future research.

**Acknowledgements** The authors would like to thank the schools, children and their guardians who helped us to facilitate this research. We would also like to thank Dr Alan Barker from Children's Health and Exercise Research Centre, University of Exeter, UK, for valuable comments during the early manuscript preparation phase.

**Contributors** LJ, IR, SÅ, HJS, UMK and THT contributed to planning this work. IR, J-PK and SÅ contributed to analyses. All authors contributed to the interpretation of the data, drafting and reporting the work, and revising critically the intellectual content. All authors have given final approval for this version, agreed to be accountable and are committed to resolving possible questions related to its content.

**Funding** This work was supported by the Juho Vainio Foundation (201410342) and the Finnish Ministry of Education and Culture (OKM/92/626/2013). IR and SÅ received funding from Business Finland and IR a grant from the Jenny and Antti Wihuri Fund.

**Competing interests** None declared.

**Patient and public involvement** Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

**Patient consent for publication** Not required.

**Ethics approval** The original study setting was approved by the ethics committee of the University of Jyväskylä. Participants and their guardians delivered a signed informed consent. All measurements were carried out by trained personnel and in accordance with the Declaration of Helsinki.

**Provenance and peer review** Not commissioned; externally peer reviewed.

**Data availability statement** Raw is agreed not to be shared with third parties. In other cases, data are available upon reasonable request. Please contact THT for data sharing.

**Open access** This is an open access article distributed in accordance with the Creative Commons Attribution Non Commercial (CC BY-NC 4.0) license, which permits others to distribute, remix, adapt, build upon this work non-commercially, and license their derivative works on different terms, provided the original work is properly cited, appropriate credit is given, any changes made indicated, and the use is non-commercial. See: <http://creativecommons.org/licenses/by-nc/4.0/>.

### ORCID iD

Laura Joensuu <http://orcid.org/0000-0002-9544-6552>

### REFERENCES

- Collins FS, Varmus H. A new initiative on precision medicine. *N Engl J Med* 2015;372:793–5.
- Ross R, Goodpaster BH, Koch LG, et al. Precision exercise medicine: understanding exercise response variability. *Br J Sports Med* 2019;53:1141–53.
- Gevaert AB, Adams V, Bahls M, et al. Towards a personalised approach in exercise-based cardiovascular rehabilitation: How can translational research help? A 'call to action' from the Section on Secondary Prevention and Cardiac Rehabilitation of the European Association of Preventive Cardiology. *Eur J Prev Cardiol* 2020;27:1369–85.
- Gluckman PD, Hanson MA. The Developmental Origins of Health and Disease. In: Wintour EM, Owens JA, eds. *Early life origins of health and disease. advances in experimental medicine and biology*. Springer, 2006.
- Veijalainen A, Tompuri T, Haapala EA, et al. Associations of cardiorespiratory fitness, physical activity, and adiposity with arterial stiffness in children. *Scand J Med Sci Sports* 2016;26:943–50.
- Faria WF, Mendonça FR, Santos GC, et al. Effects of 2 methods of combined training on cardiometabolic risk factors in adolescents: a randomized controlled trial. *Pediatr Exerc Sci* 2020;32:217–26.
- Tomkinson GR, Lang JJ, Tremblay MS, et al. International normative 20 m shuttle run values from 1 142 026 children and youth representing 50 countries. *Br J Sports Med* 2017;51:1545–54.
- Lang JJ, Belanger K, Poitras V, et al. Systematic review of the relationship between 20m shuttle run performance and health indicators among children and youth. *J Sci Med Sport* 2018;21:383–97.
- Ruotsalainen I, Renvall V, Gorbach T, et al. Aerobic fitness, but not physical activity, is associated with grey matter volume in adolescents. *Behav Brain Res* 2019;362:122–30.
- Gracia-Marco L, Vicente-Rodríguez G, Casajús JA, et al. Effect of fitness and physical activity on bone mass in adolescents: the Helena study. *Eur J Appl Physiol* 2011;111:2671–80.
- Ruiz JR, Caverro-Redondo I, Ortega FB, et al. 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. *Br J Sports Med* 2016;50:1451–8.
- Hastie T, Tibshirani R, Friedman J. *The elements of statistical learning: data mining, inference, and prediction*. Springer Science & Business Media, 2009.
- Breiman L. Random forests. *Mach Learn* 2001;45:5–32.
- Blom A, Tammelin T, Laine K, et al. Bright spots, physical activity investments that work: the Finnish schools on the move programme. *Br J Sports Med* 2018;52:820–2.
- Booth ML, Okely AD, Chey T. The reliability and validity of the physical activity questions in the who health behaviour in schoolchildren (HBSC) survey: a population study. *Br J Sports Med* 2001;35:263–7.
- Taylor SJ, Whincup PH, Hindmarsh PC, et al. Performance of a new pubertal self-assessment questionnaire: a preliminary study. *Paediatr Perinat Epidemiol* 2001;15:88–94.
- Rajala K, Kankaanpää A, Laine K, et al. Associations of subjective social status with accelerometer-based physical activity and sedentary time among adolescents. *J Sports Sci* 2019;37:123–30.
- Piko BF. Self-Perceived health among adolescents: the role of gender and psychosocial factors. *Eur J Pediatr* 2007;166:701–8.
- Waist Circumference and Waist-Hip Ratio: Report of a WHO Expert Consultation 2008.
- Opetushallitus. M toimintakyvyn seurantajärjestelmä. No title.
- Joensuu L, Sävöja H, Kallio J, et al. Objectively measured physical activity, body composition and physical fitness: cross-sectional associations in 9- to 15-year-old children. *Eur J Sport Sci* 2018;18:882–92.
- Evenson KR, Catellier DJ, Gill K, et al. Calibration of two objective measures of physical activity for children. *J Sports Sci* 2008;26:1557–65.
- Cooper AR, Goodman A, Page AS, et al. Objectively measured physical activity and sedentary time in youth: the International children's accelerometry database (ICAD). *Int J Behav Nutr Phys Act* 2015;12:113.
- Fawcett T. An introduction to ROC analysis. *Pattern Recognit Lett* 2006;27:861–74.
- Armstrong N, Van Mechelen W, Backx FJ. *Oxford Textbook of Children's Sport and Exercise Medicine*. 3rd Ed. Oxford University Press, 2017.
- Shmueli, Koppius. Predictive analytics in information systems research. *MIS Quarterly* 2011;35:553–72.
- Genuer R, Poggi J-M, Tuleau-Malot C. Variable selection using random forests. *Pattern Recognit Lett* 2010;31:2225–36.





## IV

# UTILIZING THE INTERNATIONAL CLASSIFICATION OF FUNCTIONING, DISABILITY AND HEALTH (ICF) IN FORMING A PERSONAL HEALTH INDEX

by

Ilkka Rautiainen, Lauri Parviainen, Veera Jakoaho, Sami Äyrämö, and Jukka-  
Pekka Kauppi 2023

Manuscript

# Utilizing the International Classification of Functioning, Disability and Health (ICF) in forming a personal health index

Ilkka Rautiainen<sup>\*1</sup>, Lauri Parviainen<sup>2</sup>, Veera Jakoaho<sup>2</sup>, Sami Äyrämö<sup>1</sup>, and Jukka-Pekka Kauppi<sup>1</sup>

<sup>1</sup>*Faculty of Information Technology, PO Box 35, FI-40014 University of Jyväskylä, Finland*

<sup>2</sup>*David Health Solutions Ltd., Mannerheimintie 113, FI-00280 Helsinki, Finland*

## Abstract

We propose a new model for comprehensively monitoring the health status of individuals by calculating a personal health index. The central framework of the model is the International Classification of Functioning, Disability and Health (ICF) developed by the World Health Organization. The model is capable of handling incomplete and heterogeneous data sets collected using different techniques. The health index was validated by comparing it to two self-assessed health measures provided by individuals undergoing rehabilitation. Results indicate that the model yields valid health index outcomes, suggesting that the proposed model is applicable in practice.

**Keywords**— Personal health index, ICF, Health data, Data preparation

## 1 Introduction

Reliably describing and monitoring a person's health status can greatly aid in providing appropriate treatment for each individual. Accurate information about personal health allows for better targeting of interventions, and can provide metrics

---

<sup>\*</sup>Corresponding author:

*Email address:* `ilkka.t.rautiainen@jyu.fi`

that enhance factors related to overall health instead of focusing on treating a single symptom.

A personal health index aims to condense information about the overall health of a person to a single number [1]. While health index aims to summarize the person's health by a single number, a health profile describes the health status in a set of scores [1]. The index and/or profile can then be further employed in multiple ways. For instance, a singular value can provide a quick overall status of the person. This status can be useful for healthcare professionals and for the persons themselves, aiming at systematic and understandable way of monitoring the health status of persons from a broad perspective [1]. In addition, examining further different aspects of well-being provided by a health profile is often crucial to obtain more detailed perspective of person's health.

Concisely describing a person's overall health status can be challenging. Changes in a person's functioning and health can be attributed to several reasons. For instance, the prevention, treatment and rehabilitation of physical illness, such as back pain, include also the consideration of psychological and social aspects [2]. Therefore, it is necessary to describe all the aspects affecting a person's health in a structured and standardized form, and to present this information in an easily interpretable fashion.

Global standardization of health status measurement procedures for health index construction is difficult to achieve due to different standards and practices in different countries. There can be significant variations in treatment procedures depending on the practices of different countries, clinics and therapists. For instance, clinics in different countries may use different health questionnaires (and different languages). These differences make straightforward concatenation of data sets from separate clinics impossible.

In this study, we propose using the International Classification of Functioning, Disability and Health (ICF) [3] developed by the World Health Organization as a basis for construction of a personal health index. As a comprehensive and standardized classification of functioning, disability and health, ICF covers all relevant aspects affecting personal health, making it an ideal platform for health index development. By using the ICF framework, we first convert original measurements to new variables in "ICF space" using accepted linking procedures. These variables are called ICF codes throughout the paper according to ICF terminology. This allows standardization of possibly heterogeneous data sets from different individuals into the same data space. Then, we recursively calculate the health index from an ICF tree structure using available measurements. Calculation does not require measurements from every node of the ICF tree, making it robust to missing values. On the other hand, if measurements do not cover all relevant aspects of health, the missing parts can be taken into account in future data collection. Once the overall health index value is calculated, it is also possible to investigate values for each ICF code in the tree separately, providing insights into specific sectors of functioning, disabilities and contextual factors affecting health. This way, it is

possible to obtain a comprehensive picture of individual’s health and see whether specific aspects of health need to be improved. To the best of our knowledge, this is the first study that employs the ICF framework to form a personal health index.

Multiple health indices and health profiles have been recently suggested in literature. Meijer et al. (2010) [4] introduced an internationally comparable health index that is based on functional limitations and self-reported health measures in addition to objectively measured grip strength. Kohn (2012) [5] employed multiple correspondence analysis to form a health index, giving some freedom of choice to the index user, such as the decision of what questions to include in the index domain. Poterba et al. (2013) [6] used principal component analysis for the responses to 27 health-related questions, using the first principal component as their health index. Chen et al. (2016) [7] developed a method called MyPHI that gives a personal health index (PHI) as its output. They treated the task of creating the PHI as a soft-label optimization problem with a data mining technique originally designed for complex event detection on video material. Their method can handle geriatric data with infrequencies, incompleteness and sparsity. It is also designed to give higher weight to the latest health records. Since the output of PHI is a vector of scores with each score reflecting personal health risk in a disease category, by the McDowell (2006) [1] definition, their method outputs a health profile instead of a health index. Lai et al. (2020) [8] introduced a personal health index based on a tensor decomposition method to overcome the limitations of health examination records.

Overall, in the existing research there are several elements of what we aim to achieve in this study, such as considering the differences in data by country as well as the ability to handle sparse and infrequent data with missing values. However, previously suggested health indices and profiles for the most part rely on a pre-defined set of attributes that are used to calculate the final health index. Since there are many existing methods for collecting data, for example through functioning questionnaires [9, 10] and customized tests (e.g. with equipment available at the clinic), it would be more beneficial to combine the information from different sources effectively, and then utilize it further in the health index calculation process.

The remaining part of the study proceeds as follows. Section 2 gives a brief introduction to the ICF. This is necessary preliminary information for understanding health index calculation. Section 3 describes data used in this study, including necessary conversion steps of the original data to the ICF space. Section 4 describes actual computation of a health index. Section 5 presents the results produced during the statistical analyses. The study is discussed and concluded in Section 6.

## 2 International Classification of Functioning, Disability and Health (ICF)

The overall aim of the International Classification of Functioning, Disability and Health (ICF) is to "provide a unified and standard language and framework for the description of health and health-related states". It is a complementary classification to the ICD-10/11 diagnosis classification [3].

The ICF brings together two conceptual paradigms of disability: the medical paradigm and the social paradigm. The medical paradigm sees disability as something which requires medical care, the cause being a disease, trauma or other health condition of the individual. Moreover, in the social paradigm disability is seen as a socially-created problem, something that requires a political response. The fusion of these two paradigms in the ICF can be seen as a "biopsychosocial" approach. In other words, the ICF provides a synthesis of biological, individual and social aspects of health [3, 11]. The ICF model has potential to change the current disease-based model of care, from anticipating or reacting to individual diseases to *healthy ageing* approach, which aims to observe individual's trajectories longitudinally in order to support personalized interventions proactively [12].

An overview of the structure of the ICF is presented in Fig. 1. The four health measurements/questionnaires we are utilizing and linking to the ICF are shown at the bottom of the figure, and they are not part of the ICF itself. These elements are discussed later in Section 3.1.

The ICF classifies health and health-related issues and it consists of two main parts, *Functioning and disability* and *Contextual factors*, which are further divided into four main components: *Body functions (b)* (physiological functions of body systems, including psychological functions), *Body structures (s)* (anatomical parts of the body), *Activities and participation (d)* (execution of a task or action by an individual and involvement in a life situation) and *Environmental factors (e)* (physical, social and attitudinal environment in which people live). Although *Personal factors* (background of an individual's life and living) are part of the structure, they can not currently be classified using ICF [3].

The main components have over 1,400 subcategories that are spread over four hierarchical levels. The categories in ICF are nested; the broader categories are defined to include more detailed subcategories of the parent category. For example, the chapter two ICF code, *The eye, ear and related structures (s2)*, in the *Body structures* component includes separate categories on the structures of eye socket, eyeball, around eye and so on [3].

The *chapter level* ICF code is indicated by writing a single digit number after the component *b*, *s*, *d* or *e*, for example *b2* referring to the ICF code *Sensory functions and pain*. The *second level* ICF code is specified using a three digit number (e.g. *b280* is the *Sensation of pain*). Furthermore, the *third level* ICF code has four numbers (e.g. *b2801* is the *Pain in body part*), and finally the *fourth level* has five numbers (e.g. *b28013* is the *Pain in back*) [3, 11].

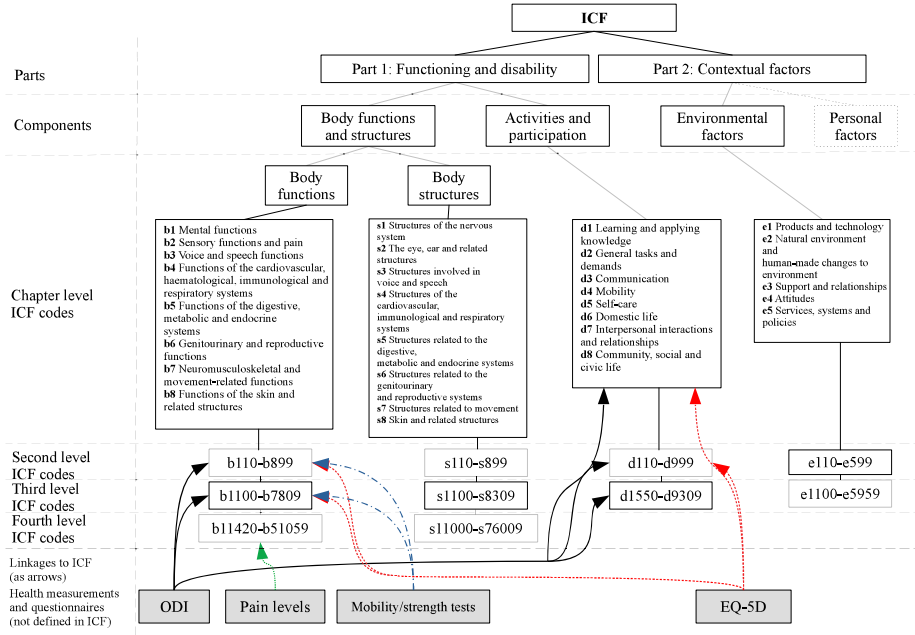


Figure 1: The hierarchical structure of the ICF (combined from the information available in [3]) with the utilized health measurements and questionnaires and their linkages to ICF.

To become a *classification* and therefore complete, the ICF code is equipped with one or more *qualifiers*. A qualifier is a single-digit number that is used to denote, for instance, the magnitude of the level of health or severity of the problem at issue. When the qualifier is written for an ICF code, a dot (.) is used as a separator between the ICF code and the qualifier(s). For example, a complete ICF code with a qualifier would be *b280.1*, which would indicate that the person is feeling a mild/slight *Sensation of pain* [3].

All of the possible ICF codes support the first qualifier, though its usage varies slightly between different main constructs. For the ICF codes under *Body functions* (*b*) and *Body structures* (*s*) the first qualifier is a *generic qualifier* that indicates the extent or magnitude of an impairment. A generic qualifier can have values 0 (**no** problem), 1 (**mild** problem), 2 (**moderate** problem), 3 (**severe** problem) or 4 (**complete** problem), with additional values 8 and 9 reserved for not specified and not applicable states, respectively. For the ICF codes under the *Activities and participation* (*d*) component the first generic qualifier indicates *performance*, a problem in the person’s current environment [3, 11].

In the case of *Environmental factors* (*e*), the generic qualifier can be either a *barrier* or a *facilitator*, meaning that a positive contributor can also be recorded. The negative scale is employed similarly to the aforementioned constructs for the barriers, while the facilitators are indicated using a plus sign (+) as an ICF

code/qualifier separator instead of a dot (e.g. ICF code  $e145 + 2$  would indicate that products for education are a moderate facilitator, as opposed to  $e145.2$ , which would mean that the products of education are a moderate barrier) [3, 11]. The facilitators are not, however, part of our proposed health index calculation model.

In addition to the first qualifier, it is possible to define additional qualifiers for *Body structures* and *Activities and participation*. For example, the nature of impairment can be described using the second qualifier in *Body structures* (e.g.  $s7300.32$  indicates a partial absence of the upper extremity) [3, 11]. However, since we utilize only the first qualifier in our proposed calculation model, these additional qualifiers are not discussed here further. Throughout the rest of this paper, the term *qualifier* will refer only to the first qualifier.

### 3 Data

For the development and validation of the health index, we employed data collected from a single David Health Solutions clinic between 2013–2019. The data included 505 persons (age  $48 \pm 18$  y, min 12 y, max 87 y, 259 females and 246 males) receiving rehabilitation treatment for various problems, including back, neck, hip, knee, shoulder, general health, and other unspecified reasons. Our original data consisted of following questionnaire and measurement data sets:

- Oswestry low back pain disability questionnaire (ODI) [9]
- a generic health questionnaire EQ-5D-5L [10]
- mobility and maximal isometric strength tests using various spine concept rehabilitation machines
- pain level answers

A required preprocessing step in the health index construction is to convert, or link, variables of these original data sets to new variables called ICF codes. Linkage of original variables to ICF codes is described next.

#### 3.1 Linkage of data to the ICF

The ICF linkage guidelines [13, 14] recommend that two medical professionals are trained to create the linkages between the original data sources to the ICF code qualifiers. In the linking process, two experts independently created the linkage information for a new data source. If the two independent linkages were identical they were accepted as is. In case they differed, a third expert opinion was considered when deciding the final linkage. Each measurement  $s, m \in \mathbb{Z}$  was always mapped linearly to the range  $0 \leq x \leq 4$ , which is the range of the qualifiers in the ICF. Details of linkage for each questionnaire and measurement data set are available in A.

Since there can be both scientifically validated and self-made ICF linkages present in the data, it is essential to be able to define separate linkage reliability values for different linkage types.  $r \in [0, 1]$  is the **linkage reliability** defined for the corresponding source. The index user must define the  $r$  value for every available source before calculating the index. For instance, we might want to define independently validated linkages with maximum reliability 1, while self-made linkages can be considered less reliable. Thus, their  $r$  would in most cases be defined to be less than 1.

For the two questionnaires, EQ-5D and ODI, data were available for 111 and 147 persons, respectively. The pain level answers were accessible for 348 persons and mobility/strength test results for 420 persons. Additionally, EQ-VAS, self-assessed health status, was answered by 168 persons.

Most often available ICF code in the dataset was *b780 (Sensations related to muscles and movement functions)*, which was available for 420 persons out of 505. The next three accessible ICF codes were *b7305 (Power of muscles of the trunk)*, *b7355 (Tone of muscles of trunk)*, and *b7401 (Endurance of muscle groups)*. They were all available for 388 persons. See Fig. 2 for a full list of available ICF codes. To be included in the list, there had to be at least one measurement available for the specified ICF code.

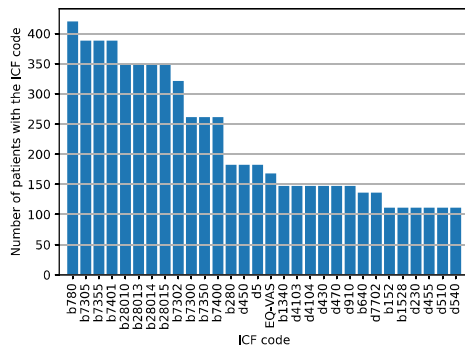


Figure 2: An overview of the occurrence of ICF codes in the David dataset for the persons. Additionally, EQ-VAS answer is also included.

### 3.2 Overview of data

There are two important properties to examine in the data. Firstly, *duration* of a treatment period is the number of full calendar days between the beginning and the end of the latest day of the treatment. By this definition, when treatment is made only once, the duration of the treatment is 0. Secondly, we can examine the frequency or number of treatment days during the treatment period. We call this property a *length* of the treatment sequence. For example, if we have a person that has been treated for three weeks and had four visits to the clinic, the duration of



the treatment period would be 21 days, with four as the length of the treatment sequence.

The median duration of the treatment period in data was 69 days, while the median length of the treatment sequence was three days. In Fig. 3 we examine these two properties further.

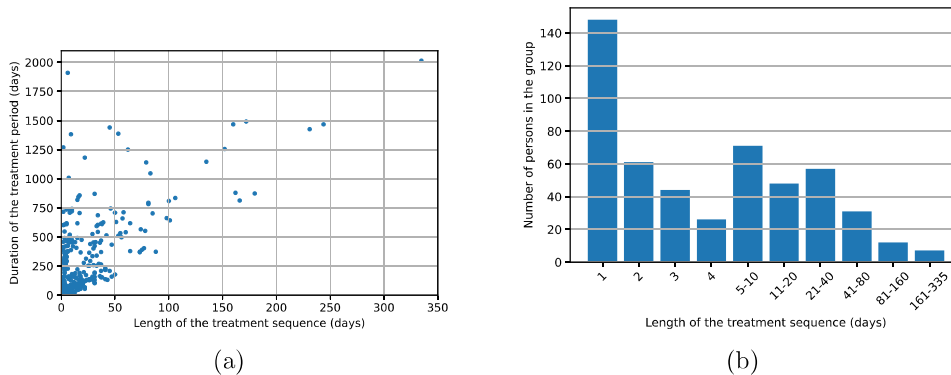


Figure 3: In (a) duration of the treatment periods are shown. The duration of the treatment period and the length of the treatment sequence are presented for each person. The majority of the persons are in the lower-left corner, meaning that they have had relatively short treatment with few individual treatment days. (b) shows how many days there is data from. Ten separate groups for the length of the treatment sequences were formed to examine how many persons were in each group. For instance, we see that for over 140 persons data is available for one day only.

## 4 Calculating the health index

In this section, methods for calculating the health index are presented. We first introduce a structure of the model used in calculation, involving its basic elements. After that, we describe overall calculation process of the health index from the model. Then, we present model elements in more detail, followed by details of the ICF code level calculations.

### 4.1 Model structure and elements

The proposed model structure corresponds to the hierarchical structure of the ICF. However, instead of using all possible ICF codes, the proposed model structure consists of available ICF codes only. For instance, ICF codes available in this study are listed in Fig. 2, so the structure of the model consists of these ICF codes only. Within this restricted set of all possible ICF codes, some ICF codes can further be missing, i.e. they are not measured for all persons. Thus, some ICF codes in

the model are observed codes whereas some codes are empty codes. Moreover, an ICF code in the model can be observed through not only one but multiple measurements. This situation happens when more than one original data variables are linked to a same ICF code. It is also possible to have multiple measurements in a single ICF code by linking one original data variable recorded from different time points. In this study, all linked measurements are called qualifiers according to the ICF terminology.

For illustration purposes, Fig. 4 gives an example of a simplified tree structure containing ICF codes. As mentioned above, multiple original measurements can be linked to a same ICF code, meaning that the corresponding ICF code is observed through multiple qualifiers. In Fig. 4 we have two ICF codes with two qualifiers linked to the same ICF code on the third and fourth levels. Firstly, the first qualifier in the ICF code  $b28010$  (4.1) is 2 (first number inside the first square brackets), with  $time\_elapsed = 0$  and  $reliability = 1$ . The second qualifier, a measurement coming from a different source, for the same ICF code is 1 (second number inside the first square brackets), with  $time\_elapsed = 30$  and  $reliability = 0.8$ . Finally, the second ICF code with two qualifiers is  $b2801$  (3.3). Also empty ICF codes are included in Fig. 4 for illustration purposes.

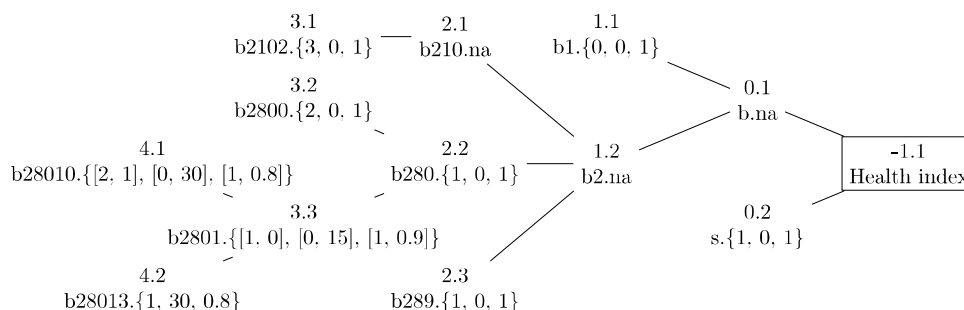


Figure 4: An example of a tree containing ICF codes. Each ICF code has information in two rows. The first row has the level number and an ICF code number in the level, separated by a dot. The format used is  $level.index$ , where  $level$  is the ICF level the node belongs to, level 4 being the deepest, and  $index$  is the per-level running index. The ICF codes are in alphabetical order at each level, and the running index is determined by this order. The level labeled as 0 refers to the ICF top components  $b$ ,  $d$ ,  $e$  and  $s$ , the level 1 to the chapter level ICF codes, the level 2 to the second level ICF codes and so on. The second ICF code row indicates the ICF code and three attributes associated with it, ICF qualifier  $x$ , time elapsed from measurement ( $TE$ ) (see B.2) and reliability of the linkage  $r$  (see Section 3.1). The format is  $ICF\_code.\{qualifier, time\_elapsed, reliability\}$ . When a qualifier is not available for the ICF code,  $na$  is shown instead after the dot.

## 4.2 Health index calculation

The principal idea in the proposed approach is to transport information node-by-node and level-by-level from the deepest level to the root of the tree, eventually resulting in a single number called the health index. The calculation procedure for the health index is started from the leaf nodes of the tree. For example, in Fig. 4 the calculation starts from ICF code b28010 (4.1), then continuing on to b28013 (4.2) and then to the next level ICF code b2102 (3.1) etc. The calculation procedure is repeated through the tree to the right until there are no ICF codes left. The last calculated value is the unscaled health index, and scaling this value to 0–100 produces the final health index. The calculation procedure is described in Alg. 1. The equations referred to in the algorithm are presented after the algorithm.

---

### Algorithm 1: Health index calculation.

---

**Data:** ICF code tree with qualifiers  
**Result:** Health index

```

1 initialize variable level with the deepest ICF code level in the tree
2 while the health index has not been calculated do
3   index = 1
4   while there are ICF codes at the level do
5     if there are child ICF codes for the ICF code level.index then
6       calculate value of the ICF code  $x_{level.index}$  (Eqs. 2, 3, 4)
7       calculate time weighting  $\alpha_{level.index}$  (Eq. 5)
8       calculate linkage reliability  $r_{level.index}$  (Eq. 5)
9       foreach child ICF code of level.index do
10        | remove all qualifiers s if they exist
11        | index = index + 1
12   level = level - 1
13   if level < -1 then
14     | scaled_health_index =  $HI(x_{-1,1})$  (Eq. 1)
15     | return scaled_health_index

```

---

In Alg. 1 the input data is the ICF code tree with qualifiers, similar to the one depicted in Fig. 4. The output result is the health index. In the beginning (line number 1), we initialize the variable *level*, according to the deepest ICF code level available in the data. For example, in the data depicted in Fig. 4, there are qualifiers present in two ICF codes at level 4 (4.1 and 4.2). Thus, the *level* variable in this case would be initialized as 4. The lowest possible *level* is -1, corresponding to the root of the tree. At this level, the final health index is calculated.

The purpose of the outer **while** loop (Alg. 1, lines 2–15) is to process ICF codes of the tree one by one until the root of the tree has been reached and the

health index calculated and returned. Inside the loop, the variable *index* is first initialized (line 3). The *index* acts as a per-level running index, i.e. it is reset to 1 every time the level of the tree changes. For example, the processing order of the ICF codes in the tree shown in Fig. 4 is 4.1, 4.2, 3.1, 3.2, 3.3, and so on.

The inner **while** loop (Alg. 1, lines 4–11) serves a purpose of repeating the necessary calculations until all the ICF codes in the current *level* are handled. The first **if** statement (line 5) checks whether there are child ICF codes available for the currently examined ICF code, denoted as *level.index*. For example, in Fig. 4 the child ICF codes of *b280* are *b2800* and *b2801*.

In case child ICF codes are available for an ICF code, calculations can start. In lines 6–8 the value of the ICF code,  $x_{level.index}$ , time weighting  $\alpha_{level.index}$  and linkage reliability  $r_{level.index}$  are calculated. These calculations are discussed in detail in Sec. 4.4.

In lines 9–10 every qualifier is removed from the child ICF codes of *level.index*. The purpose of this procedure is to make sure these values are not used later in the calculations. Instead, the values calculated during lines 6–8 are employed when the calculation continues further.

The per-level running *index* is increased after each ICF code has been handled (Alg. 1, line 11), and the level index *level* decreased after all ICF codes in the level have been handled (line 12).

When the level index *level* goes below  $-1$ , it means that the root of the tree has been reached (Alg. 1, line 13). The raw health index, denoted  $x_{-1.1}$  by the ICF code numbering scheme employed in Fig. 4, has been calculated. The raw value will be finally transformed to make it more decipherable. A suitable target presentation for the final health index is an integer number from 0 (worst health) to 100 (best health). To achieve this, the source and target values are inverted: while raw health index value 0 indicates best possible health, the same value in the target range indicates worst possible health.

The transformation is applied in the following way, with the final health index denoted as *HI* (Alg. 1, line 14):

$$HI(x_{-1.1}) = \text{nint} \left( 100 - 100 \times \frac{x_{-1.1} - \min(x_{-1.1})}{\max(x_{-1.1}) - \min(x_{-1.1})} \right), \quad (1)$$

where *nint* is the nearest integer function and  $x_{-1.1}$  is the raw value to be transformed, while  $\min(x_{-1.1})$  and  $\max(x_{-1.1})$  are the minimum and maximum theoretical or actual raw values. Since we know the theoretical minimum (0) and maximum (4) values of the raw health index, our first option is to directly scale and invert the raw values to our target. In this case we would define that  $\min(x_{-1.1}) = 0$  and  $\max(x_{-1.1}) = 4$ . The returned value after the transformation (Alg. 1, line 15) is the final health index.

After the health index has been calculated, the tree is reset to its former state, meaning that all calculated values are removed from the ICF codes, and all qualifiers are restored to their original positions. This procedure's meaning is to

ensure that any future calculations do not use any previously calculated values.

### 4.3 A closer look at the elements

In Fig. 5 we take a closer look at the elements in the tree. Here we have taken three ICF codes,  $b28010$ ,  $b28013$  and  $b2801$ , from the Fig. 4 for a more thorough inspection. An ICF code, whose output value is calculated, is denoted by  $q$ . In Fig. 5, this ICF code is equivalent to node 3.3 ( $b2801$ ) shown in Fig. 4. Furthermore,  $q$ 's child ICF codes,  $ch1_q$  and  $ch2_q$ , are equivalent to ICF codes 4.1 ( $b28010$ ) and 4.2 ( $b28013$ ), respectively.

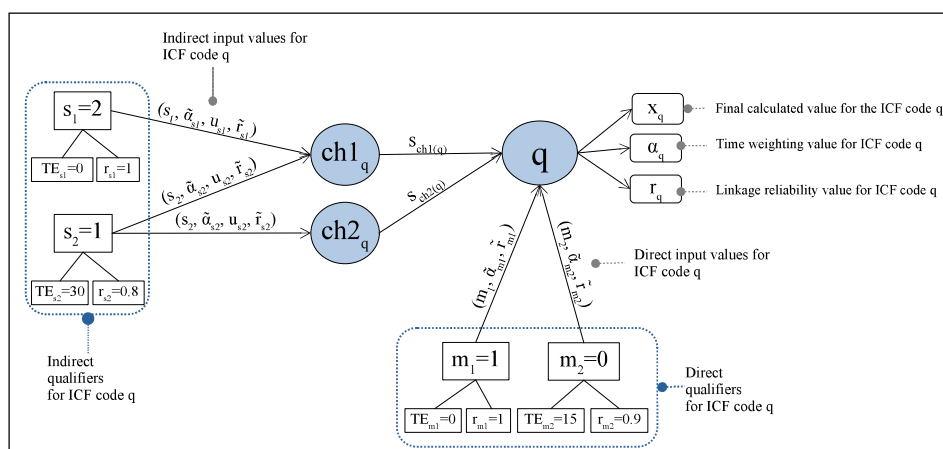


Figure 5: An ICF code level example of the health index calculation. In this figure we have a viewpoint of an individual ICF code marked as  $q$ . Its child ICF codes are marked here as  $ch1_q$  and  $ch2_q$ .

In this hypothetical sample data there are four separate qualifiers, indicated with markings  $s_1$ ,  $s_2$ ,  $m_1$  and  $m_2$ , with linkages between qualifiers and ICF codes indicated with arrows:

1. A *moderate problem* (qualifier  $s_1 = 2$ ) is recorded for ICF code *Pain in head and neck* ( $b28010/ch1_q$ ). The measurement is made in the latest available day in the treatment period, as indicated by  $TE_{s_1} = 0$  and its linkage reliability is defined as the strongest possible ( $r_{s_1} = 1$ ).
2. A *slight problem* (qualifier  $s_2 = 1$ ) is recorded for ICF codes *Pain in head and neck* ( $b28010/ch2_q$ ) and *Pain in back* ( $b28013/ch2_q$ ). The measurement is made 30 days ago from the perspective of latest available measurement, as indicated by  $TE_{s_2} = 30$ , and its linkage reliability is defined as ( $r_{s_2} = 0.8$ ).
3. A *slight problem* (qualifier  $m_1 = 1$ ) is recorded for ICF code *Pain in body part* ( $b2801/q$ ). The measurement is made in the latest available day in the

treatment period, as indicated by  $TE_{m_1} = 0$  and its linkage reliability is defined as the strongest possible ( $r_{m_1} = 1$ ).

4. A *no problem* (qualifier  $m_2 = 0$ ) is recorded for ICF code *Pain in body part* ( $b2801/q$ ). The measurement is made 15 days ago from the perspective of latest available measurement, as indicated by  $TE_{m_2} = 15$ , and its linkage reliability is defined as ( $r_{m_2} = 0.9$ ).

The different elements are marked in Fig. 5:

- **Direct qualifiers (D) for ICF code q:**  $m_1$  and  $m_2$  are the **direct** qualifiers for ICF code  $q$ , i.e. they are directly linked to  $q$ .  $TE_{m_1}$  is the age of a qualifier in full days and  $r_{m_1}$  is the reliability of the linkage.
- **Direct input values for ICF code q:** These values ( $m_i, \tilde{\alpha}_{mi}, \tilde{r}_{mi}$ ) are attached to the ICF code  $q$ . While  $m_i$ , the qualifier, can be directly attached to the ICF code, other values need to be calculated based on the available data.
- **Indirect qualifiers (I) for ICF code q:** This element depicts the two available qualifiers  $s_1$  and  $s_2$  for the two child ICF codes  $ch1_q$  and  $ch2_q$ .  $s_1$  and  $s_2$  are the qualifiers for child ICF codes of ICF code  $q$ , i.e. they are the **indirect** qualifiers for ICF code  $q$ . The indirect qualifiers are linked to  $q$  through its child ICF codes.  $TE_{s_1}$  is the time elapsed value and  $r_{s_1}$  is the reliability of the source linkage.
- **Indirect input values for ICF code q:** These values ( $s_j, \tilde{\alpha}_{sj}, u_{sj}, \tilde{r}_{sj}$ ) are attached to the child ICF codes, and are used to calculate the values,  $s_{ch1(q)}$  and  $s_{ch2(q)}$ , for the child ICF codes.

There are three outputs in the diagram:

- **Final calculated value for the ICF code q:**  $x_q$  is the final value for ICF code  $q$ . It consists of **direct** qualifiers and **indirect** qualifiers made for the child ICF codes.  $x_q$  is the weighted average of all available qualifiers:  $s_1, s_2, m_1$  and  $m_2$ . In this case it would be calculated as follows:  $x_q = (\tilde{\beta}_{m_1}^D m_1 + \tilde{\beta}_{m_2}^D m_2) + s_{ch1(q)} + s_{ch2(q)} = (\tilde{\beta}_{m_1}^D m_1 + \tilde{\beta}_{m_2}^D m_2) + (\tilde{\beta}_{s_1}^I s_1 + \tilde{\beta}_{s_2}^I s_2) + (\tilde{\beta}_{s_2}^I s_2)$ , where  $\beta$  terms refer to weighting coefficients. Details of the equation are explained in Section 4.4.
- **Time weighting value for q:**  $\alpha_q$  is the weighted mean of all available time weighting  $\alpha$  values defined in Eq. 5.
- **Linkage reliability value for q:**  $r_q$  is the weighted mean of all available  $r$  values, defined also in Eq. 5.

## 4.4 Calculation of ICF code and reliability values

In the previous section, we explained computation of the ICF code value through an example for better understanding. In this section, we describe calculation of ICF code and reliability values more generally.

### 4.4.1 ICF code value calculation

The value of the ICF code  $q$  is defined as:

$$\begin{aligned} x_q &= f\left(\sum_{i \in D} \tilde{\beta}_{mi}^D m_i + \sum_{k \in ch_q} s_{chk(q)} + \sum_{k \in ch_q} x_{chk(q)}\right) \\ &= f\left(\sum_{i \in D} \tilde{\beta}_{mi}^D m_i + \sum_{k \in ch_q} \sum_{j \in k} \tilde{\beta}_{sj}^I s_j + \sum_{k \in ch_q} \tilde{\beta}_{xk}^I x_k\right). \end{aligned} \quad (2)$$

This definition consists of four separate elements:

1. Function  $f$  is a weighting function that can be selected to either give emphasis to higher qualifiers (exponential weighting) in data or to highlight lower qualifiers (logarithmic weighting). It is also possible to use no weighting at all (linear weighting). The different weighting functions are presented in more detail in B.1.
2. The first sum term  $\sum_{i \in D} \tilde{\beta}_{mi}^D m_i$  is the weighted sum of qualifiers  $m_i$  that are directly linked with the ICF code  $q$ . In the equation,  $D$  refers to a set of **direct qualifiers**, i.e. those qualifiers that are directly linked to  $q$ . In other words, they are the direct qualifiers from the viewpoint of  $q$ . Additionally, we define a total weighting term  $\beta^D$  for direct qualifiers:

$$\beta_{mi}^D = \tilde{\alpha}_{mi} \tilde{r}_{mi}. \quad (3)$$

As can be seen, the total weighting consists of two elements  $\alpha$  and  $r$ . The first element  $\alpha$  is a time weighting value that depends on age of the corresponding qualifier. Generally, as a qualifier gets older, the less weight it is given in the calculation. Time weighting is discussed in more detail in B.2. The second element  $r$  is the linkage reliability that we defined in Section 3.1. It should be noted that  $\alpha$ ,  $r$  and  $\beta$  in Eq. 2 carry the tilde notation ( $\tilde{\alpha}$ ,  $\tilde{r}$ ,  $\tilde{\beta}$ ). This notation means that a normalized value for the term is used instead of the raw value. A standard normalization procedure is carried out to make the different values directly comparable. The two normalization equations are presented in B.3.

3. The second sum term  $\sum_{k \in ch_q} \sum_{j \in k} \tilde{\beta}_{sj}^I s_j$  is the weighted sum of qualifiers  $s_j$  that are indirectly, i.e. via the child ICF codes of  $q$ , linked with the ICF code  $q$ . In the outer sum,  $ch_q$  refers to a child ICF code of  $q$  that has a

qualifier or a calculated value. Thus, each child ICF code of  $q$  is utilized in the calculation. The inner sum runs through all the qualifiers linked to the child ICF code. A total weighting term for indirect qualifiers is called  $\beta^I$ . It is a slightly modified version from the weighting term shown in Eq. 3, defined as:

$$\beta_{sj}^I = \tilde{\alpha}_{sj} \tilde{r}_{sj} u_{sj}, \quad (4)$$

where  $u_{sj}$  is the uniqueness of source of the qualifier  $s_j$ , calculated based on how many child ICF codes the same qualifier is linked to. It is defined as  $u_{sj} = 1/z_j$ ,  $z_j \in \mathbb{N}^+$ , where  $z_j$  is the number of linkages between  $s_j$  and child ICF codes of  $q$ .  $u$  is designed to lower the qualifier weighting if the same source is used to form values for two or more ICF codes under the same parent ICF code.  $\alpha$  and  $r$  have the same purpose as in Eq. 3.

4. The third sum term  $\sum_{k \in ch_q} \tilde{\beta}_{xk}^I x_k$  is similar to the second sum term above. The difference is that it runs through the calculated values ( $x$ ), instead of the qualifiers ( $s$ ). Here the second sum term is not needed, as each child ICF code can have only one calculated value. In addition, uniqueness of source is not applicable here. Hence,  $\tilde{\beta}_{xk}^I$  is defined similarly to Eq. 3, as  $\tilde{\beta}_{xk}^I = \tilde{\alpha}_{xk} \tilde{r}_{xk}$ .

#### 4.4.2 ICF code reliability calculations

There are two values,  $\alpha_q$  (equivalent of  $\alpha_{level.index}$  on Alg. 1, line 7, the weighted mean of time weighting  $\alpha$ ) and  $r_q$  (equivalent of  $r_{level.index}$  on line 8, the weighted mean of linkage reliability  $r$ ), that are calculated for  $x_q$ . They are defined almost identically to  $x_q$  in Eq. 2:

$$\alpha_q = \sum_{i \in D} \tilde{\beta}_{mi}^D \alpha_{mi} + \sum_{k \in ch_q} \sum_{j \in k} \tilde{\beta}_{sj}^I \alpha_{sj} + \sum_{k \in ch_q} \tilde{\beta}_{xk}^I \alpha_{xk}, \quad (5)$$

and similarly for  $r_q$ , by replacing the  $\alpha$  term in the equation with  $r$ . Here the  $\tilde{\beta}$  terms are used to determine the weightings for each  $\alpha/r$  element. These values,  $\alpha_q$  and  $r_q$ , "travel" with the  $x_q$  value and are associated with it.

## 5 Statistical analyses

The set of calculated health indices was validated by comparing the results with person's self-reported EQ-VAS answers and maximum pain level answers. Two groups were formed for validation. For the first group the treatment period duration was at least 90 days, whereas for the second group the same limit was 30 days. Length of treatment sequence (see Section 3.2) in the first group was at least 10 days. For the 30-day group, the length of treatment sequence was at least



five days. A person could be assigned to both groups, provided that the day and qualifier preconditions were satisfied.

To examine the effect of different time weightings, Pearson correlation coefficients were calculated for different time decay constants  $\gamma$  (see B.2 for detailed definition of  $\gamma$ ). The values for  $\gamma$  were chosen so that:

- $\gamma_1 = (1/20)^{1/30} \approx 0.905$  represented a very heavy time decay: a 30 day old qualifier was weighed only 5 % of its original time weight.
- $\gamma_2 = (1/3)^{1/30} \approx 0.964$  represented our estimated preliminarily potential time decay value in a real-life scenario. In practice this value means that a measurement made 30 days ago was weighed as one third of its original time weight.
- $\gamma_3 = 1$  means that there was no time decay at all.

The algorithm was implemented using Python programming language [15]. The statistical analyses were made using SciPy [16] and figures with Matplotlib [17].

## 5.1 EQ-VAS answer vs. the health index

First, the health index was calculated without giving any emphasis to lower or higher qualifiers. In Table 1 the EQ-VAS answer was compared with the health index, and Pearson correlations are presented for both of the groups. The influence of changing the  $\gamma$  was also explored. The results show mostly moderate positive correlations for all configurations between the EQ-VAS answer and the calculated health index. For the 90-day group there were 84 persons, who had in total 125 EQ-VAS answers recorded. Most persons had only one EQ-VAS answer recorded, but some had two, three or even four answers. All of these answers were used in the calculations. The 30-day group had 115 persons with 159 EQ-VAS answers.

$\gamma$	30-day group	90-day group
$(\frac{1}{20})^{1/30}$	0.690***	0.642***
$(\frac{1}{3})^{1/30}$	0.700***	0.659***
1	0.654***	0.599***

Table 1: Pearson correlations for the EQVAS answer vs. Health index. Significance of the correlation coefficients is indicated by stars: \*\*\* indicates  $p < 0.001$ .

## 5.2 Maximum pain vs. the health index

In the case of maximum pain versus the health index, the Pearson correlation was calculated individually for each person between the maximum pain trajectory

and the health index trajectory. The maximum pain is defined as the single highest value in the four pain level answers. A maximum pain trajectory was then formed for each person. Since this information was also used in calculating the health index, there was inevitably at least some correlation between the two values. However, this comparison provides a useful second perspective to the validation of the model.

Bonferroni correction is a method designed to prevent the data from incorrectly showing significance. The method is applied here in the maximum pain vs. health index trajectory correlation calculations, and  $\alpha$  level is divided by  $n$ .

Distribution of correlation values for the two groups are presented in Fig. 6. A slight upward trend in median correlation is visible in both groups: as  $\gamma$  increases, the median correlation gets closer to zero. In all cases, the correlation was negative, varying roughly from low to moderate. There are small differences in the  $n$  values for the different  $\gamma$  groups. In some configurations, the health index value remained constant for some persons, and thus calculating the correlation was not possible. These persons were omitted from the calculations.

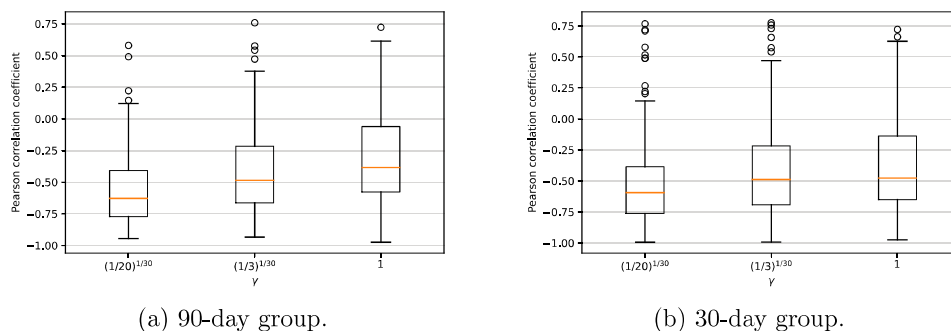


Figure 6: Boxplots of Pearson correlations between the health index and maximum pain trajectories. (a) shows the 90-day group, where correlation medians for the three time decay groups ( $n = 133$ ) are -0.628, -0.486, and -0.385, respectively. (b) shows the 30-day group, where correlation medians for the three time decay groups are -0.594, -0.489 ( $n = 184$ ), and -0.477 ( $n = 182$ ).

Looking at the portions of significant correlations, for the 90-day group, Bonferroni corrected significant (\*) portions were 55 % for  $\gamma_1$ , 41 % for  $\gamma_2$  and 31 % for  $\gamma_3$ . In the 30-day group, the similar portions were 41 % for  $\gamma_1$ , 29 % for  $\gamma_2$  and 23 % for  $\gamma_3$ . Thus, there are more significant correlations when length of the treatment sequence and duration of the treatment period increase.

In Table 2 the effect of time series length is further examined by binning the maximum pain vs. health index trajectories of the persons into three bins based on the length of treatment sequence. These results also indicate that as more data points are included in the trajectory, significant portion increases.

Bin	Day ranges	Bonferroni significant (*) portions	Median correlations	Subset lengths
1	[10, 26], [5, 15]	18.8 %, 7.9 %	-0.385, -0.569	48, 63
2	[27, 42], [16, 32]	36.6 %, 18.3 %	-0.500, -0.376	41, 60
3	[43, 325], [33, 325]	70.5 %, 65.6 %	-0.510, -0.511	44, 61

Table 2: Maximum pain vs. health index correlation statistics when using three bins with the length of treatment sequence as the binning criterion. The three bins are roughly the same size (tertiles). Here the time decay value was defined as  $\gamma_2$ . Both, the 90-day group and the 30-day group, are presented, separated by a comma. *Day ranges* shows, for each bin, the lengths of treatment sequences included. For example, the *Bin 1* includes persons with length of treatment sequence from 10 to 26 days (the 90-day group) or persons with length of treatment sequence from 5 to 15 days (the 30-day group). *Bonferroni significant (\*) portions* refers to the portion of Bonferroni corrected significant correlations at level 0.05. *Median correlations* shows the median of all correlations in the bin, while *Subset lengths* gives the number of persons used for each bin.

In Fig. 7 two example persons and their maximum pain vs. health index trajectories are presented. Firstly, there is a typical example with moderate negative correlation between the trajectories. Secondly, trajectories showing very high negative correlation are presented.

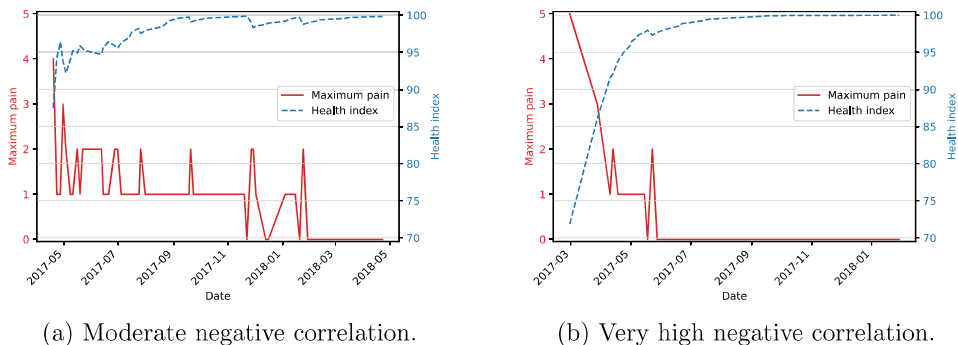


Figure 7: Two cases of correlations between the maximum pain and health index trajectories. (a) is a typical case showing moderate negative correlation (-0.670) with a length of treatment sequence 40 days and (b) is an extreme case showing very high negative correlation (-0.934) with a length of treatment sequence 28 days. Time decay was defined as  $\gamma_2$  in both cases. Bonferroni corrected correlations were significant (\*\*\*)

### 5.3 Effect of non-linear weighting

In this section, the effect of qualifier weighting for correlations is examined. Qualifier weighting can be calibrated in the model by tuning the  $y$  parameter. There

are three different functions that the weighting can use: 1) when  $y \in (0, 2)$ , the weighting is exponential, meaning that higher qualifier values have more weight, 2) when  $y \in (2, 4)$ , the weighting is logarithmic; lower qualifier values have more weight, and 3) when  $y = 2$ , the weighting is linear; qualifier values are treated without any weighting. Details on the value weighting can be found in B.1.

For the  $y$  parameter, 15 evenly spaced values were chosen for both sides of the linear ( $y = 2$ ) case, meaning in total  $15 + 1 + 15 = 31$   $y$  parameter values were examined from  $y = 0.2$  to  $y = 3.8$  with intervals of 0.2. The results for health index versus EQ-VAS answer using Pearson correlation are shown in Fig. 8. For the 90-day group, highest correlations for all three  $\gamma$  values were observed at  $y = 2.12$ . The correlations were 0.643, 0.664, and 0.599, respectively. In the 30-day group the highest correlations were observed at the linear weighting point ( $y = 2$ ). Correlations at that point are the same presented in Table 1.

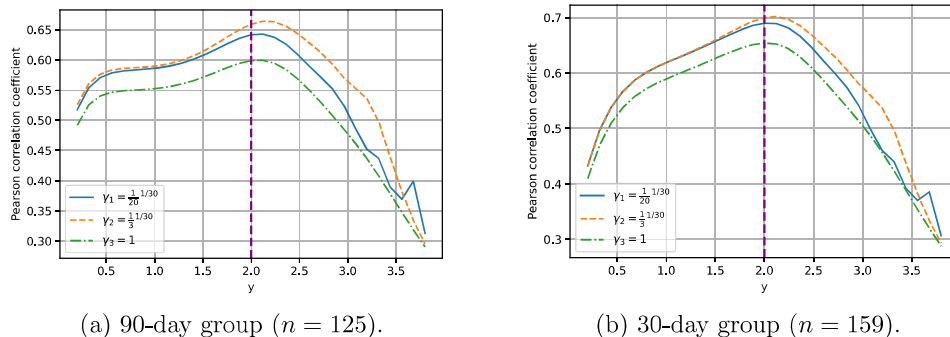


Figure 8: Pearson correlation between the health index and EQ-VAS answer.

In case of maximum pain vs. the health index, median correlations for the two groups are presented in Fig. 9. Here, the highest negative correlations observed for the three  $\gamma$  values for the 90-day group were -0.644 (when  $y = 2.24$ ), -0.491 and -0.420 (both when  $y = 2.36$ ), respectively. In the 30-day group the correlations were -0.618 (when  $y = 2.36$ ), -0.489 and -0.477 (both when  $y = 2$ ).

## 6 Discussion

### 6.1 Conclusions

The ICF framework provides an ideal platform for developing a health index, as it is a widely-used, standardized classification system that covers comprehensively aspects affecting person's health and functioning. Its validity has been proven by numerous published studies and it is heavily utilized worldwide for population health research projects [18]. Another advantage of using the ICF is that

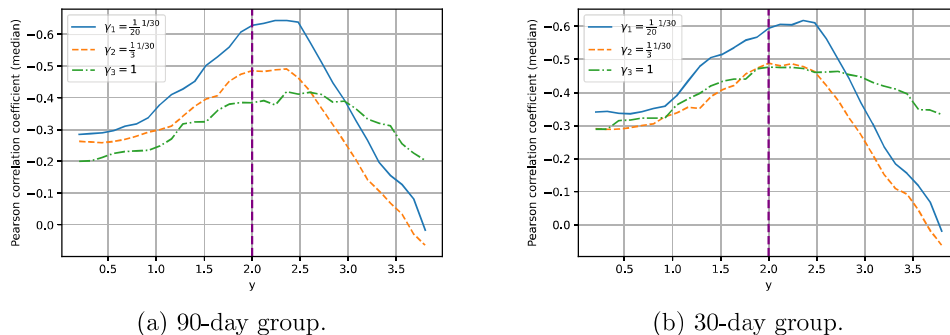


Figure 9: Median Pearson correlation between the health index and maximum pain. (a) shows the 90-day group ( $\gamma_1, \gamma_2: n = 133, \gamma_3: n = 132, 133$ ) and (b) the 30-day group ( $\gamma_1, \gamma_2: n = 184, \gamma_3: n = 181, 182, 183$ ).

many questionnaires used to evaluate treatment response have been linked to ICF codes, allowing questionnaire results to be converted to ICF code qualifiers. These linkages have been scientifically validated and there are extensive guides [19, 14] available for creating new linkages for newer datasets.

The proposed method can also be easily applied to calculate a health profile. In the context of the ICF, a health profile would most likely be a set of separate scores for each of the four ICF *components* (a. Body functions and structures, b. Activities and participation, c. Environmental factors, d. Personal factors). It is also possible to create a more detailed profile by going deeper into the ICF structure.

This study established a framework for homogenization of data sets across clinics so that they can be concatenated into a single very large data set. This step is highly crucial for development of AI, because training of machine learning algorithms requires that same variables are available from all analyzed persons. If homogenization of data sets is not performed, only subsets of persons and/or variables can be analyzed at once, reducing predictive power of the trained models. When going a few steps further, one of the goals is to employ the health index as an optimization target in a machine learning model. This approach will make it possible to improve the person’s health comprehensively instead of focusing on a single health parameter and can, for example, help in choosing the optimal rehabilitation pathway for a person.

In addition to being able to handle heterogenous data from various international sources, it is expected that the proposed index can be formed even when there are only few qualifiers available for a person. However, when there are more qualifiers available, the health index becomes more reliable. One of the aims was also to emphasize recent qualifiers in health index calculation, while also giving some weight to qualifiers from the past.

## 6.2 Limitations

The ICF framework enables its users to define health and functioning of an individual in very detailed ways. In *Body structures (s)* construct the second qualifier can be used to indicate the nature of the change in the respective body structure, and the third qualifier indicates the location, e.g. left or right [3, 19]. Furthermore, in *Activities and participation (d)* the second generic qualifier (capacity) is used to indicate limitation without assistance [3]. The proposed model utilizes only the first generic qualifier. Thus, any further information potentially provided by these additional qualifiers is not utilized in the current model.

In the case of ICF's *Environmental factors (e)* construct the first qualifier can be used to denote either the positive effects (facilitators) or the negative effects (barriers) of the environment [3]. The model is currently able to handle only barriers. Thus, the potential positive effect of possible facilitators to the health index is not part of the proposed model.

Qualifiers 8 (not specified) and 9 (not applicable) are not handled by the proposed model. Currently the values are treated as ordinal, and these nominal values can not be utilized. This limitation might result in loss of some information, since these qualifiers are used in ICF system to record information that can not be captured using the  $[0, 4]$  value range. Qualifier 8 should be applied when there is a problem, but it is unknown whether that problem is mild or severe. Additionally, qualifier 9 is typically used in a situation when use of the category is not appropriate for the individual [19].

A potential limitation regarding the practical usability of the model is that too scarce data might give a wrong impression about person's overall health. The proposed model does not currently specify any metrics for observing how complete the employed data is. Any personal factors that are not expressed by the qualifier format, such as age of a person, can not be utilized in the current calculation process.

## CRediT authorship contribution statement

**Ilkka Rautiainen:** Methodology, Software, Validation, Formal analysis, Investigation, Writing - Original Draft, Visualization, **Lauri Parviainen:** Conceptualization, Methodology, Data Curation, **Veera Jakoaho:** Data Curation, Writing - Review & Editing, Project administration, **Sami Äyrämö:** Conceptualization, Methodology, Writing - Review & Editing, Supervision, **Jukka-Pekka Kauppi:** Conceptualization, Methodology, Software, Writing - Review & Editing, Visualization, Supervision

## Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## Acknowledgements

Ilkka Rautiainen's work was funded by David Health Solutions Ltd. and Jenny and Antti Wihuri Foundation (grant numbers 00180312, 00200302, and 00210295).

## References

- [1] I. McDowell, *Measuring health: A guide to rating scales and questionnaires*, Oxford University Press, 2006.
- [2] G. A. Fava, N. Sonino, Psychosomatic medicine: Emerging trends and perspectives, *Psychotherapy and psychosomatics* 69 (4) (2000) 184–197.
- [3] World Health Organization, *International Classification of Functioning, Disability and Health: ICF*, World Health Organization, Geneva, 2001.
- [4] E. Meijer, A. Kapteyn, T. Andreyeva, Internationally comparable health indices, *Health Economics* 20 (5) (2011) 600–619. doi:10.1002/hec.1620.
- [5] J. L. Kohn, What is health? A multiple correspondence health index, *Eastern Economic Journal* 38 (2) (2012) 223–250. doi:10.1057/eej.2011.5.
- [6] J. M. Poterba, S. F. Venti, D. A. Wise, Health, education, and the post-retirement evolution of household assets, Working Paper 18695, National Bureau of Economic Research (January 2013). doi:10.3386/w18695.
- [7] L. Chen, X. Li, Y. Yang, H. Kurniawati, Q. Z. Sheng, H.-Y. Hu, N. Huang, Personal health indexing based on medical examinations: A data mining approach, *Decision Support Systems* 81 (2016) 54–65. doi:10.1016/j.dss.2015.10.008.
- [8] G. Lai, D. Yu, S. Zhang, Z. Wei, X. Sun, Personal health index based on residential health examination, in: X. Yang, C.-D. Wang, M. S. Islam, Z. Zhang (Eds.), *Advanced Data Mining and Applications*, Springer International Publishing, 2020, pp. 569–581. doi:10.1007/978-3-030-65390-3\_43.
- [9] J. Fairbank, J. Couper, J. Davies, J. O'Brien, et al., The Oswestry low back pain disability questionnaire, *Physiotherapy* 66 (8) (1980) 271–273.

- [10] M. Herdman, C. Gudex, A. Lloyd, M. Janssen, P. Kind, D. Parkin, G. Bonsel, X. Badia, Development and preliminary testing of the new five-level version of EQ-5D (EQ-5D-5L), *Quality of Life Research* 20 (10) (2011) 1727–1736. doi:10.1007/s11136-011-9903-x.
- [11] World Health Organization, *Towards a common language for functioning, disability, and health: The International Classification of Functioning, Disability, and Health* (2002).
- [12] M. Cesari, I. Araujo de Carvalho, J. Amuthavalli Thiyagarajan, C. Cooper, F. C. Martin, J.-Y. Reginster, B. Vellas, J. R. Beard, Evidence for the domains supporting the construct of intrinsic capacity, *The Journals of Gerontology: Series A* 73 (12) (2018) 1653–1660. doi:10.1093/gerona/gly011.
- [13] A. Cieza, T. Brockow, T. Ewert, E. Amman, B. Kollerits, S. Chatterji, T. B. Ustün, G. Stucki, Linking health-status measurements to the International Classification of Functioning, Disability and Health, *Journal of Rehabilitation Medicine* 34 (5) (2002) 205–210. doi:10.1080/165019702760279189.
- [14] A. Cieza, N. Fayed, J. Bickenbach, B. Prodinger, Refinements of the ICF linking rules to strengthen their potential for establishing comparability of health information, *Disability and Rehabilitation* 41 (5) (2019) 574–583. doi:10.3109/09638288.2016.1145258.
- [15] G. Van Rossum, F. L. Drake, *Python 3 Reference Manual*, CreateSpace, Scotts Valley, CA, 2009.
- [16] P. Virtanen, R. Gommers, T. E. Oliphant, M. Haberland, T. Reddy, D. Cournapeau, E. Burovski, P. Peterson, W. Weckesser, J. Bright, et al., SciPy 1.0: fundamental algorithms for scientific computing in Python, *Nature Methods* 17 (3) (2020) 261–272. doi:10.1038/s41592-019-0686-2.
- [17] J. D. Hunter, Matplotlib: A 2D graphics environment, *Computing in Science & Engineering* 9 (03) (2007) 90–95. doi:10.1109/MCSE.2007.55.
- [18] R. H. Madden, A. Bundy, The ICF has made a difference to functioning and disability measurement and statistics, *Disability and Rehabilitation* 41 (12) (2019) 1450–1462. doi:10.1080/09638288.2018.1431812.
- [19] World Health Organization, *How to use the ICF: A practical manual for using the International Classification of Functioning, Disability and Health (ICF), Exposure draft for comment*. Geneva: WHO 10 (2013).
- [20] M. Koç, B. Bayar, K. Bayar, A comparison of back pain functional scale with Roland Morris disability questionnaire, Oswestry disability index, and short form 36-health survey, *Spine* 43 (12) (2018) 877–882.



- [21] R. D. Baruah, P. Angelov, D. Baruah, Dynamically evolving clustering for data streams, in: 2014 IEEE Conference on Evolving and Adaptive Intelligent Systems (EAIS), 2014, pp. 1–6. doi:10.1109/EAIS.2014.6867473.

## Appendix A Details on data linkages

The Oswestry low back pain disability questionnaire (ODI) includes one item on pain and nine items on activities of daily living (lifting, walking, social life, personal care, sitting, standing, sleeping, traveling, and sex life), each scored on a 0–5 scale, 5 representing the highest disability [20]. We defined the linkages ourselves, and they are depicted in Table A.1. Further, the translations between the original ODI answer and the equivalent value as an ICF code qualifier are shown in Table A.2. The original ODI answer refers to the selected response number.

Item as appeared	Purpose of information in question	ICF code(s)
Pain intensity	Level of pain	b280
Personal care (washing, dressing etc.)	Pain related to personal care tasks	b280, d5
Lifting	Pain related to lifting objects	b280, d430
Walking	Pain related to walking	b280, d450
Sitting	Pain related to sitting	b280, d4103
Standing	Pain related to standing	b280, d4104
Sleeping	Pain related to sleep	b280, b1340
Sex life (if applicable)	Pain related to sexual activities	b280, d7702, b640
Social life	Pain related to participation in social activities	b280, d910
Travelling	Pain related to capacity for travelling	b280, d470

Table A.1: The ODI questions, their purpose, and their equivalent ICF codes. The response options are not listed.

Original ODI answer:	0	1	2	3	4	5
ICF code qualifier:	0	1	2	3	4	

Table A.2: The original ODI answers and their translated equivalent values in the ICF system.

The five-level version of the EQ-5D generic health questionnaire, EQ-5D-5L consists of five questions on mobility, self-care, usual activities, pain/discomfort, and anxiety/depression. The scale of answers is on a 1–5 scale, with 5 being the highest disability. In addition to the five questions, there is a self-assessed score for

overall health on a 0–100 scale, EQ-VAS. It is not used when calculating the health index. Instead, EQ-VAS is only used for the external validation of the index. Since, similarly to the 0–4 scale of the ICF qualifier, there are five response options in all the questions excluding the EQ-VAS. Thus, the answers can be directly mapped as ICF code qualifiers. The questions and their linkages to ICF codes are described in Table A.3. We defined these linkages ourselves.

<b>Item as appeared</b>	<b>Purpose of information in question</b>	<b>ICF code(s)</b>
Mobility	Problems with walking	d450, d455
Self-care	Problems with self-care, specifically washing and dressing	d5, d510, d540
Usual activities (e.g. work, study, housework, family or leisure activities)	Problems with daily routine / usual tasks	d230
Pain / Discomfort	Level of pain or discomfort	b280
Anxiety / Depression	Level of anxiety or depression	b152, b1528
We would like to know how good or bad your health is TODAY.	General health perception	N/A

Table A.3: The EQ-5D-5L questions, their purpose, and their equivalent ICF codes. The response options are not listed.

Pain levels in the back, hip/leg, neck, and shoulder/arm areas were measured from persons in the visual analogue scale (VAS) on a 0–10 scale in the beginning of every visit to a clinic. Because these variables were measured frequently, it is important to include these data sets in the analysis besides questionnaire data, which was collected typically only once or twice during a whole treatment period. We mapped all the four pain level answers ourselves into ICF.

In Table A.4 the linkages between the original different pain level questions and the ICF codes are shown. Further, the translations between the original pain level answer and the equivalent value as an ICF code qualifier are shown in Table A.5. For example, the original pain level answer of "3" for back pain would translate to ICF code *b28013* as qualifier "1".

We also created new linkages for the mobility/strength tests. Strength and mobility levels related to relevant spine functions were measured using David Health Solutions' spine concept rehabilitation machines and presented in newton-metres and degrees, respectively. These values were converted to relative changes with respect to the average values of a reference population. Better mobility/strength than reference was mapped to 0 %, which corresponded to *no problem*. The interventions employed with the machines as well as their ICF codes are described

Pain level	Corresponding ICF code
Back	b28013 (Pain in back)
Hip/leg	b28015 (Pain in lower limb)
Neck	b28010 (Pain in head and neck)
Shoulder/arm	b28014 (Pain in upper limb)

Table A.4: The original pain level questions and their equivalent ICF codes.

Original pain level:	0	1	2	3	4	5	6	7	8	9	10
ICF code qualifier:	0		1		2			3		4	

Table A.5: The original pain level answers and their translated equivalent values in the ICF system.

in Table A.6. The translations of the original relative change values to the ICF code qualifiers are shown in Table A.7. The values were then converted to the ICF domain employing the linkages depicted using the information in these two tables. For example, the relative change of 20 % observed through the f120 machine would be linked with qualifier "1" to the ICF codes *b780*, *b7305*, *b7355* and *b7401*.

Intervention	Target of intervention	Corresponding ICF codes
110 Trunk Extension	Increase of dorsal, lumbar and thoracic region muscle tone and strength, Pain prevention	b7305 (Power of muscles of the trunk) b7355 (Tone of muscles of trunk) b7401 (Endurance of muscle groups) b780 (Sensations related to muscles and movement functions)
130 Trunk Flexion	Increase of abdominal region muscle tone and strength, Pain prevention	
120 Trunk Rotation	Increase of lateral and abdominal region muscle tone and strength, Pain prevention	b7302 (Power of muscles of one side of the body) b7305 (Power of muscles of the trunk) b7355 (Tone of muscles of trunk) b7401 (Endurance of muscle groups) b780 (Sensations related to muscles and movement functions)
150 Trunk Lateral Flexion	Increase of lateral and abdominal region muscle tone and strength, Pain prevention	
140 Cervical Extension / Lateral Flexion	Increase of cervical region muscle tone and strength, Pain prevention	b7300 (Power of isolated muscles and muscle groups) b7302 (Power of muscles of one side of the body) b7350 (Tone of isolated muscles and muscle groups)
160 Cervical Rotation	Increase of cervical region muscle tone and strength, Pain prevention	b7400 (Endurance of isolated muscles) b780 (Sensations related to muscles and movement functions)

Table A.6: The original spine concept rehabilitation machine interventions, their purpose and their equivalent ICF codes.

Original relative change (%):	$0 \leq x \leq 4$	$4 < x \leq 24$	$24 < x \leq 49$	$49 < x \leq 95$	$95 < x \leq 100$
ICF code qualifier:	0	1	2	3	4

Table A.7: The original spine concept rehabilitation machine relative changes (%) and their translated equivalent values in the ICF system.

## Appendix B Defining weightings and normalization

There are two separate weighting elements in the health index: time decay weighting and different value weighting functions. The purpose of the time weighting is to give more emphasis to newer qualifiers. By using the value weighting functions there is a possibility to tune measured values, either to give more emphasis to higher handicap levels or to downplay their role in the data.

### B.1 Value weighting functions

The weighting function is selected by first defining a tuning parameter  $y \in ]0, 4[$  that determines the steepness of the weighting function. The curve always starts from point  $(0, 0)$ , is fitted to go through point  $(2, y)$  and always ends in point  $(4, 4)$ . The selected  $y$  value directly effects the function type used in fitting. There are three types of functions.

The function is exponential when  $y \in (0, 2)$ , defined as:

$$f(x) = ae^{bx} + c. \quad (\text{B.1})$$

The function is logarithmic when  $y \in (2, 4)$ , defined as:

$$f(x) = a \ln(bx + 1). \quad (\text{B.2})$$

Finally, the function is linear when  $y = 2$ . Linear function is simply defined as  $f(x) = x$ .

The values for  $a$ ,  $b$  and  $c$  in exponential and logarithmic functions are solved during the standard curve fitting process. For all weighting functions apply  $x \in [0, 4]$ , meaning the range is the same as in the generic qualifier in the ICF. The three weighting functions are visualized in Fig. B.1, using values  $y = 0.75$  for the exponential and  $y = 3.25$  for the logarithm function. Furthermore, selecting exponential weighting function will mean that higher values, i.e. higher handicap levels, in data are emphasized. By contrast, when logarithmic weighting is employed, all input values are increased by the function, meaning that lower handicap levels are emphasized.

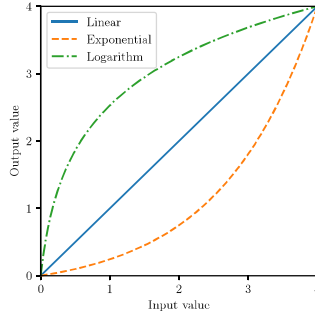


Figure B.1: Weighting functions visualized using example  $y$  values.

## B.2 Time weighting

After linking the data to the ICF, the next step is to form a table that for each individual person contains all ICF code qualifiers available. In order to prepare the time series of different persons more easily comparable, the date of the person's any first qualifier is always defined as 0. Each consecutive measurement day for the person is then given a number  $d \in \mathbb{N}$ , expressing in full calendar days the distance from the start of the treatment. We can then define  $TE$ , the time elapsed in days from the latest valid qualifier:

$$TE = d - d_0, \quad (\text{B.3})$$

where  $d_0$  is the date of the person's latest valid qualifier, excluding  $d$ . Raw time weighting  $\alpha$  is then defined as:

$$\alpha = \gamma^{TE}, \quad (\text{B.4})$$

where  $\gamma \in (0, 1]$  is the time decay constant defined by the user before calculating the index. Therefore, lower  $\gamma$  indicates a stronger time decay. This approach was inspired by [21].

## B.3 Normalization of the time weighting, linkage reliability and $\beta$ term

To make sure that the different time weightings and reliability values are comparable, we can calculate the normalized time-weighting values  $\tilde{\alpha}_{mi}$  and  $\tilde{\alpha}_{sj}$ , normalized reliability values  $\tilde{r}_{mi}$  and  $\tilde{r}_{sj}$  as well as normalized  $\beta$  term values  $\tilde{\beta}_{mi}$  and  $\tilde{\beta}_{sj}$  as follows:

$$\tilde{\alpha}_{mi} = \frac{\alpha_{mi}}{\sum_{i \in D} \alpha_{mi} + \sum_{k \in ch_q} \sum_{j \in k} \alpha_{sj} + \sum_{k \in ch_q} \alpha_{xk}} \quad (\text{B.5})$$

and

$$\tilde{\alpha}_{sj} = \frac{\alpha_{sj}}{\sum_{i \in D} \alpha_{mi} + \sum_{k \in ch_q} \sum_{j \in k} \alpha_{sj} + \sum_{k \in ch_q} \alpha_{xk}}, \quad (\text{B.6})$$

where  $\alpha$  is the raw non-normalized time-weighting that we defined in Eq. B.4. Additionally, the normalized reliability values and normalized  $\beta$  term values are calculated identically, by replacing the  $\alpha$  terms in equations with  $r$  or  $\beta$ , respectively. As a result of this normalization, the sum of all normalized terms is equal to one.