UNIVERSITÀ DEGLI STUDI DI MILANO

PhD course in Nutritional sciences
XXXIV cycle

Machine learning applied to clinical nutrition

Andrea Foppiani

Supervisor: Prof. Alberto Battezzati
Doctoral Programmes Coordinator: Prof. Luciano Pinotti

2021

Abstract

Introduction. The healthcare sector is failing to utilize routinely produced clinical data to refine the care experience and to augment knowledge created in clinical study. The big data culture and the closely connected field of machine learning constitute the latest and the best opportunity yet to put to good use data created as a by-product of clinical care. Aim of this thesis was to test the capabilities of machine learning algorithms applied to real-world clinical nutritional data to assist clinicians in their decision making process. Machine learning was used in two predictive contexts: 1) prediction of routinely collected parameters for patients non-eligible for the reference method, and 2) prediction of failure to meet clinical targets set out for the patient.

Material and methods. A large nutritional dataset collected at the International Center of Nutritional Status (University of Milan, Milan, Italy) was used for the analysis. The dataset include 15780 patients and multi-domains predictors providing informations on age, sex, education, occupation, marital status, family status, menstruation, pregnancies, diet status, diet history, physical activity, smoking, pharmacological treatments, clinical signs, weight history, physical exam, blood pressure, anthropometry, bioimpedance analysis, ultrasound (abdomen fat thicknesses), indirect calorimetry, laboratory exams, anxiety, depression, binge eating, emotion regulation, eating disorders, and adherence to a Mediterranean diet.
Machine learning algorithms were applied in order to predict the following outcomes: resting energy expenditure by indirect calorimetry, total body water by bio-impedance analysis, weight loss failure, failure to improve basal glycemia, failure to improve total cholesterolemia, failure to improve triglyceridemia. To evaluate accuracy and discrimination ability of machine learning and statistical algorithms, a series of cross-validation experiments were conducted for all outcomes, and the most accurate algorithm for each outcome was selected as the best for that outcome. Accuracy was defined with the root-mean-square error for continuous outcomes and the correct classification fraction for categorical outcomes.

Results. Machine learning algorithms outperformed statistical algorithms for all outcomes. The best performing models were tree-based models, in particular bagged decision trees performed best for continuous outcomes, while random forests performed best for categorical outcomes (with the exception of the triglyceridemia outcome which saw a boosted tree algorithm as the best performer).
In the prediction of resting energy expenditure and total body water, accuracy was high and the mean errors were deemed small in the context of clinical practice [mean (95% confidence interval) root-mean-square error 27.6 (20.9, 34.3) kcal/day and 0.842 (0.768, 0.916) l respectively].
In the prediction of weight loss failure, failure to improve basal glycemia, failure to improve total cholesterolemia, and failure to improve triglyceridemia, the mean correct classification fraction ranged between .616 and .735, but even the best algorithms showed good sensitivity but poor specificity (mean area under the ROC curve ranged between .652 and .687). For categorical outcomes unbalanced toward the event, machine learning models were the only one able to improve the accuracy of a naive

classifier that assumes that all patients will experience the event, although only in weight loss failure model outcome accuracy was consistently above the naive classifier.

Discussion. Our results highlight the ability of machine learning algorithms to provide a high-accuracy alternative to reference techniques for non-eligible patients. The big-data culture paired with machine learning algorithms seem able to overcome limitations imposed from using externally-developed equations, providing highly accurate predictions.
In the setting of identifying non-responders, machine learning algorithms did not provide highly discriminant predictions, but were the only ones able to provide a better prediction of random guessing or the historical rate of event. In this more ambitious task, machine learning algorithm results need to be critically interpreted by the clinician, whose reasoning is necessarily different but can incorporate the suggestions provided from these algorithms.

Introduction

A global learning health system

THE HEALTHCARE SECTOR might be expected to be relentless in seeking to identify and apply what is best in practice, but a mixture of factors works in combination to generate a sector that is frequently performing below its potential. Some of these factors are socio-cultural (innate caution rooted in a principle of non nocere, vested interests, health system complexity, and established professional norms), while others are informational (the evidence required to guide best decisions is either not available, or if available, not accessed or used by the decision-makers).

EVIDENCE-BASED PRACTICE is the movement embodying this emphasis on clinical evidence to guide practice. However, evidence based practice is an ideal that, for many reasons, is difficult to achieve:

Knowledge availability is imperfect, as knowledge may either be missing, be fuzzy in nature, or be difficult to access by the physician.

Knowledge update should be sought by clinicians throughout their career. Nevertheless, it was estimated that each individual evidence base has an estimated half-life of 7 years. Learning takes valuable time and new paradigms of care may seem alien or threatening to clinicians.

Knowledge creation is itself flawed. The scientific community have elaborated rigorous study design and statistical analysis to produce evidence that enables market entry of new therapeutic strategies, but little is known about longer term outcomes, about how new solutions are put into practice outside the rigorous conditions in which they are tested, about interaction with other conditions or treatments, or about drift in practice. These factors contribute to the so-called reproducibility crisis, according to which many scientific studies, even the most rigorous ones, are impossible to replicate outside the initial conditions.

ISLANDS OF INFORMATION constitute the ground on which the health sector operates. Information seldom has the means to escape those islands, to enable rapid learning from experience, and put that learning efficiently to work. It reportedly takes 17 years before a new element of validated clinical knowledge finds its way into routine clinical practice in the United States.

This is in sharp contrast to other sectors, such as industry and commerce. Both those sectors maximise the use of its internal data and resultant information, embodying a learning organisation philosophy. This approach overrides any local parochialisms of practice, or indeed, when safety is concerned, it also overrides commercial competition. This surely must provide a vision and goal for the health sector to maximise effectiveness through data, as the key to safety and efficiency. Health outcomes and health status should be recognised as the measure of system performance, and data from a range of inputs should be continuously analysed to produce operational intelligence, to enable learning of what works and what doesn't.

Of course, there are major ethical, cultural, and political differences between health and the other sectors that have deployed learning models. These differences should not be used as a reason why learning health systems cannot be created. Rather, they are factors that must be taken into account in building a learning system that can promote individual and population health while conserving increasingly scarce national and global resources.

A LEARNING HEALTH SYSTEM is increasingly taking shape. In the United States the Institute of Medicine (Institute of Medicine 2011), a long-time proponent of the concept, defines a learning health system as:

> "…one in which progress in science, informatics, and care culture align to generate new knowledge as an ongoing, natural by-product of the care experience, and seamlessly refine and deliver best practices for continuous improvement in health and health care."

Big data and artificial intelligence

THE ERA OF BIG DATA has begun over recent years thanks to the digitalization, collection and storage of vast quantities of data in combination with advances in data science (Docherty and Lone 2015). The potential benefits of these large databases to health are significant, with faster progress in improving health, better value for money and higher quality science. Also, this may be of particular benefit to clinical care and research.

There is no specific definition for big data, but it is generally understood to refer to datasets whose size, complexity and dynamic nature are beyond the scope of traditional data collection and analysis methods. The analysis of such complex data requires methods more familiar to the field of informatics than clinical research, such as machine learning and computational linguistics.

Generally, big data in clinical research comes in 2 forms: data on a huge number of variables per subject or data on a huge number of subjects. Collecting a huge number of variables, more so from different fields, is important because diseases do not occur in isolation: they result from an interaction between genetic, molecular, environmental, and lifestyle factors. On the other hand, to find meaningful associations between variables, a huge number of subjects needs to be studied.

ARTIFICIAL INTELLIGENCE (AI) is a part of computer science that tries to make computers more intelligent (Kononenko 2001). One of the basic requirements for any intelligent behavior is learning and most of the researchers today agree that there is no intelligence without learning. Therefore, machine learning is one of the major branches of AI and, indeed, it is one of the most rapidly developing subfields of AI research.

Machine learning is the scientific discipline that focuses on how computers learn from data. It arises at the intersection of statistics, which seeks to learn relationships from data, and computer science, with its emphasis on efficient computing algorithms. This marriage between mathematics and computer science is driven by the unique computational challenges of building statistical models from massive data sets (Deo 2015). Two types of learning exist, supervised learning and unsupervised learning.

Supervised learning starts with the goal of predicting a known output or target. Supervised learning focuses on classification, which involves choosing among subgroups

to best describe a new instance of data, and regression, which involves estimating an unknown continuous parameter. Supervised learning is often used to estimate risk. In modeling risk, the computer is doing more than merely approximating physician skills but finding novel relationships not readily apparent to human beings.

In *unsupervised learning* there are no outputs to predict. Instead, we are trying to find naturally occurring patterns or groupings within the data. This is inherently a more challenging task to judge and often the value of such groups learned through unsupervised learning is evaluated by its performance in subsequent supervised learning tasks.

Exploiting big data for clinical care and clinical research

MEDICAL RESEARCH is currently not meeting the information needs of patients, clinicians, administrators, and policy makers. The flow of new knowledge is too slow, and its scope is too narrow. The medical research community's delay in adopting big-data approaches has left it particularly ill prepared for a precision medicine future that is designed to provide personalized information and individualized care. Medicine aspires to be a learning health care system, but it is failing to rapidly learn through data generated by individuals in clinical care and daily life. Big data and machine learning can facilitate efforts in both clinical care and clinical research.

First, data-driven thinking and methods can play a critical role in the emergence of personalized healthcare (Chawla and Davis 2013). Numerous diseases have preventable risk factors or at least indicators of risk. However, the possible combination of risk factors is so complex, it's impossible for an individual physician to fully analyze it in real time at the time of patient interaction. Currently, providers take careful histories and do physical examinations and selective laboratory testing to determine patient health and risk for future disease. These are generally limited to a few diseases and by the skill and knowledge of the individual provider and competing priorities for individual visits. Thus, taking the next big steps in personalized healthcare requires a computing and analytics framework to aggregate and integrate all the information collected by clinicians and the main patient outcomes. This would allow us to discover deep knowledge about patient similarities and connections, and provide personalized disease risk profiles for each individual patient, derived from not only the electronic medical record information of that patient, but also from similarities of that patient to other patients.

Second, medical practice and clinical research are still largely anchored in producing new knowledge through studies that tend to narrow the research question and avoid the complexities of real-world practice. Clinical trials, for instance, often exclude complicated patients (eg. those who may have several medical ailments and complex treatment regimens), but those patients are the patients typically seen in medical offices. These studies are most commonly focused on a single question, are often expensive, and take years to complete. Moreover, most studies are poorly equipped to explore how various factors may interact to influence the result for a particular patient. Meanwhile, data generated every day, for a variety of practical purposes, could serve as a practically inexhaustible source of knowledge to fuel a learning healthcare system.

However, to date, these data are largely wasted as a source of research and rarely investigated.

There are at least two ways in which machine learning can aid clinicians and medical researchers:

- *clinical decision support system,* using supervised learning to link data collected in real time to future outcomes
- *new patterns recognition,* using unsupervised learning to represent high dimensional data to find naturally occurring and possibly novel patterns on which to base further investigations

Both will be explored in the following chapters.

Clinical decision support systems

Clinical decision support systems (CDSSs) are usually embedded within electronic health records and go from pop-up alerts to more sophisticated tools incorporating clinical prediction rules or models. In the latest review of such systems, Kwan et al. (2020) found on average a 5.8% improvement in proportion of patients receiving desired care, although the upper quartile of improvement was much higher (10-62%). The reviewed CDSSs weren't necessarily using AI or were based on machine learning algorithms.

THE FIRST WAVE OF AI in medicine happened in the 1970's. Testament of that are projects such as Shortliffe's work with MYCIN, Kulikowski's individualized clinical decision models, and de Dombal's computer-aided diagnosis of acute abdominal pain which incorporated newer statistical reasoning methods (probabilistic reasoning and neural networks).

The need for rigorous evaluation of the quality and impact of AI was recognised in the 1980s. While the first publications focused mainly on methodologies for evaluating performance in a laboratory environment, later papers addressed field-testing in clinical settings, to examine effects on the process of care delivery.

Later, following implementations of AI in clinical practice, studies shifted to the clinical impacts of AI and on the related methodological challenges to find adequate clinical endpoints. Building upon evidence about the benefit of AI in medicine, research then focused on reviewing the impact of AI on patient outcomes in inpatient settings, in psychiatry, and medication safety.

The challenges recognised in the early days of applying AI were among others:

- the legal and ethical issues
- capturing the context, including informal information that is not documented in the medical record, but is nevertheless part of a doctor's mental image about the patient
- the transferability of algorithms from one setting to another both with respect to patient groups and clinical setting
- the inability of AI to reason at the boundaries or outside its own application range

- capturing the dynamic nature of professional knowledge development in health care

Montani and Striani (2019) analysed the latest literature contributions to CDSSs (years 2017-2018), focusing on approaches that adopted AI techniques. Their specific goal was to understand if "classical" knowledge-based CDSSs were still being proposed, or if there has been a shift in favor of big data analytics and machine learning approaches. They found that 49⁄75 .65 papers presented data-driven CDSSs, 26⁄75 .35 presented knowledge-based CDSSs (with 6⁄26 presenting a hybrid approach). They emphasize that the need for transparency and explainability is nowadays being recognized as a central theme to be addressed by AI research. To this point, they see in the hybrid approaches a promising strategy to deal with transparency and explainability issues, combining formalized knowledge and learnt knowledge in order to improve CDSSs competence, flexibility, and, of course, explainability.

To this date, CDSSs are not widely used in the clinical setting, despite 4 decades of research showing diagnostic accuracy that rivals the performance of expert clinician. Shortliffe and Sepúlveda (2018) provide their viewpoint on this issue, establishing these points of criticality:

- black boxes are unacceptable
- time is a scarce resource
- complexity and lack of usability thwart use
- relevance and insight are essential
- delivery of knowledge and information must be respectful to the clinician
- scientific foundation must be strong

Evaluation of a CDSS needs to be carried out from its inception, to the actual use in clinical settings, and beyond:

Design and development of AI: historically, evaluation of AI was limited to the design and development phase, as implementation and use of AI systems in routine clinical practice was rare. During design and development, evaluation concentrates upon the performance of the algorithms in terms of discrimination, accuracy, and precision. Depending on the use case, one performance measure might be more important than the other.

Selection and use of AI: widespread availability of clinical data, easy to use AI development environments and online communities have resulted in rapidly growing numbers of algorithms that have become available to clinicians. When multiple algorithms are available and one must be selected, it is important to evaluate any risks of data quality issues, and poor fit of the foundational data to a new situation, such as different population and morbidity patterns. We are also interested in the decision-making performance of humans with and without AI assistance. Once an algorithm is developed, clinical validation of its utility is needed. The algorithm may be correct but is it operationally meaningful and useful? Does it fit the clinical workflow? Does it still represent up-to-date clinical knowledge? Will it change clinical decisions? What level of confidence can be given? A significant concern here is that when humans are assisted by CDSSs, they tend to over-rely and delegate full responsibility to the CDSS rather than continuing to be vigilant. This is known as automation bias and can have

dangerous consequences when the CDSS is wrong or fails, or the presenting problem is subtly unique.

Ultimately, a CDSS need to provide evidence of:

- safety and reliability, through processes able to identify and deal with predictable errors, and a monitoring system able to identify near-misses or similar problems
- quality of the advice provided
- stability of its knowledge base

ONGOING SURVEILLANCE OF AI should measure the impact of AI on patient outcomes, the experience of receiving and providing care, as well as both organisational and social impacts (Magrabi et al. 2019). Over time, the context, treatment possibilities, and patient population might change. Therefore, once implemented, ongoing surveillance is needed to monitor and recalibrate AI algorithms.

Several indicators need to be monitored through the use of AI in clinical settings:

- *system quality:* once implemented, AI system requires ongoing surveillance based on a set of measures to recalibrate AI algorithms
- *information quality:* it covers quality of data used as input for the AI as well as quality of the output information
- *service quality:* it refers to help desk-type support available for users as well as long-term feedback from users for both immediate updates as well as for the entire system development
- *system use:* it refers to utilization (e.g. use/ non-use, frequency of use) of the system output
- *user satisfaction/acceptance:* it refers to user perceptions about system output
- *outcomes:* it refer to the individual and organizational impacts

Inductive insight from clinical data

INTEGRATING CLINICAL DATA for accurate diagnosis and treatment is one of the greatest challenges in medicine. New-knowledge discovery in high-dimensional data would require novel data-driven approaches rather than more conventional hypothesis-driven approaches (Jafari et al. 2020).

This has historically been the case for genomic information. Raw genomic data is vastly beyond human capabilities of understanding, and interactions with molecular and phenotypic features should also be considered to obtain a more systematic characterization. A systematic integration of all of this available information may provide a promising approach to knowledge discovery.

CLUSTERING has been commonly used to identify subpopulations of patients with distinctive genetic variants or gene expression profiles (Jafari et al. 2020). Implementing clustering methods in the context of precision medicine is not only applicable to omics data, but also to physiological data. Also, imaging data, as a major part of health records of individuals, is now commonly utilized. Furthermore, it has been shown that utilizing biomedical annotations can potentially improve clustering analysis to obtain more biologically relevant disease categories.

Exploring subclasses of diseases and drugs is a prevalent task in medicine. While identifying the heterogeneity of patients is critical, identifying the driving characteristics of such heterogeneity shall create new paths to knowledge discovery. To understand the underlying factors that are shared by patients with similar diseases, a vast multi-domain data generating process should be employed. By introducing machine learning algorithms, such complex domains of information can be systematically evaluated, and cluster analysis may further help infer the distinctive disease patterns.

Obesity is a complex problem

High multi-domain complexity is the common denominator to several models of obesity's causality, risk factors and comorbidities. According to the biomedical perspective, obesity is essentially the result of an energy imbalance driven by individual behavior wherein energy intake exceeds energy expenditure over time. The past few decades have seen a shift toward a socio-ecological view of obesity, in which individual behavior is situated within a broader social context (Frood et al. 2013).

In representing the "obesity system" the Foresight group has found in 2007 more than 100 socio-ecological variables and more than 300 interconnections from which obesity can arise (Vandenbroeck, Goossens, and Clemens 2007). These variables come from several domains such as physiology, food consumption, physical activity, psychology, environment and social domain. Today the system would be even more complex due to progress in obesity research in the last decade.

Not only the determinants of obesity are multi-factorial, multi-domain and complex in their interactions, even the resulting condition can present itself in many different ways and can be associated with a wide array of comorbidities or with none (eg. metabolically healthy obesity) (Stefan et al. 2013).

The only way to understand obesity is as a consequence of complex interactions between many variables. It may be so complex that traditional statistical modeling (i.e. parametric models) may produce questionable conclusions (Breiman 2001).

The International Center for the Assessment of Nutritional Status (ICANS) is a nutritional care clinical facility in the Milan urban setting devoted to multidisciplinary treatment of nutritional and nutrition-related diseases. The ICANS is equipped with a unique setting of state-of-the-art instruments to assess body composition (dual-x ray energy absorptiometry, air displacement plethysmography), visceral fat (ultrasonography), body water (bioimpedance analyser), resting energy expenditure (indirect calorimetry) and biochemical parameters of interest in nutritional and metabolic studies (mass spectrometer systems, integrated analyzer for clinical chemistry and immunoassay testing).

Relevant to this project, the ICANS has collected data since its inception in an electronic health record system. The amount of data collected can be classified as big data in both size and complexity on a multi-domain level. More than 20.000 patients have been recorded, with and more than 250 data entries have been routinely recorded in the database, encompassing several domains (data from body composition, metabolic rate, habitual diet, physical activity, clinical history, physiological and behavioral, social and

demographic). Several subset of patients also received additional examinations, further increasing variables recorded.

Aims

The general aim of this thesis was to test the capabilities of machine learning algorithms to serve as the fundamental building block for a CDSS when applied to real-world clinical nutritional data.

Two possible CDSS features will be explored:

- *compensatory system for missing variables:* prediction of routinely collected parameters for patients non-eligible for the reference method (supervised learning of continuous, cross-sectional variables)
- *alert system aimed to identify non-responders:* prediction of failure to meet clinical targets set out for the patient (supervised learning of categorical, prospective variables)

In both settings, alternative models aided by unsupervised learning of the many predictors included in the dataset were fitted and possible predictive improvements tested.

Material and methods

Dataset

DATA used in the analysis was collected at the ICANS (University of Milan, Milan, Italy), as part of a large ongoing open-cohort nutritional study. Patients were recruited between January 2009 and August 2019. At baseline, patients received a full nutritional assessment, based on the assessment an hypocaloric diet was provided, and a follow-up examination was scheduled. At follow-up, patients were interviewed by a registered dietitian, anthropometric measurements were collected, and, based on baseline clinical findings, secondary endpoints were evaluated.

PATIENTS included in this study were self-referring patients seeking a weight loss program, mainly resident in Milan or nearby cities. Eligibility criteria were: age $\geq$ 18 years; not pregnant and not nursing; no condition severely limiting movements and physical activity; no severe cardiovascular, neurological, endocrine, or psychiatric disorder; prescribed an hypocaloric diet, with macro- and micronutrient levels set accordingly to the Italian recommended daily allowances (Società Italiana di Nutrizione Umana 2014), and with a Mediterranean pattern. Characteristics of the sample are presented in Table 1.

Table 1: Patient characteristics

Characteristic	**Overall**, N = 15,780	**Female**, N = 11,253	**Male**, N = 4,527
Age	47 (37, 56)	47 (36, 56)	47 (37, 57)
Education			
Bachelor	5,672 (36%)	3,988 (36%)	1,684 (37%)
Lower secondary	1,868 (12%)	1,288 (12%)	580 (13%)
Other	394 (2.5%)	296 (2.7%)	98 (2.2%)
Primary	555 (3.5%)	356 (3.2%)	199 (4.4%)
Tertiary	190 (1.2%)	155 (1.4%)	35 (0.8%)
Upper secondary	6,975 (45%)	5,079 (46%)	1,896 (42%)
Unknown	*126*	*91*	*35*
Occupation			
Freelancer	1,265 (8.1%)	711 (6.4%)	554 (12%)
Homemaker	843 (5.4%)	843 (7.6%)	0 (0%)
Laborer	489 (3.1%)	294 (2.6%)	195 (4.3%)
Office	7,119 (46%)	5,254 (47%)	1,865 (42%)
Other	2,590 (17%)	1,693 (15%)	897 (20%)
Retired	1,565 (10%)	1,101 (9.9%)	464 (10%)
Student	1,324 (8.5%)	930 (8.4%)	394 (8.8%)
Unemployed	422 (2.7%)	303 (2.7%)	119 (2.7%)
Unknown	*163*	*124*	*39*
Marital status			
Divorced	1,032 (6.6%)	779 (7.0%)	253 (5.6%)
Married	7,893 (51%)	5,480 (49%)	2,413 (54%)
Single	6,289 (40%)	4,515 (41%)	1,774 (40%)
Widowed	402 (2.6%)	357 (3.2%)	45 (1.0%)
Unknown	*164*	*122*	*42*
Physical activity level			
None	8,147 (60%)	6,015 (61%)	2,132 (57%)
<2 h/week	2,800 (21%)	2,089 (21%)	711 (19%)
2-4h /week	1,896 (14%)	1,303 (13%)	593 (16%)
4-7 h/week	625 (4.6%)	395 (4.0%)	230 (6.1%)
>7 h/week	162 (1.2%)	83 (0.8%)	79 (2.1%)
Unknown	*2,150*	*1,368*	*782*
Smoking status			
Never smoked	8,317 (53%)	6,344 (56%)	1,973 (46%)
Ex-smoker	3,262 (21%)	2,076 (18%)	1,186 (28%)
Smoker	3,967 (26%)	2,832 (25%)	1,135 (26%)
Unknown	*234*	*1*	*233*
BMI category			
Underweight	238 (1.5%)	156 (1.4%)	82 (1.9%)
Normal weight	3,483 (22%)	2,970 (26%)	513 (12%)
Overweight	5,790 (37%)	4,162 (37%)	1,628 (37%)
Obese (Class I)	3,862 (25%)	2,426 (22%)	1,436 (32%)
Obese (Class II)	1,549 (9.9%)	1,017 (9.0%)	532 (12%)
Obese (Class III)	736 (4.7%)	507 (4.5%)	229 (5.2%)
Unknown	*122*	*15*	*107*
High fasting glucose	734 (4.7%)	376 (3.3%)	358 (7.9%)
High total cholesterol	2,041 (13%)	1,438 (13%)	603 (13%)
High triglycerides	622 (3.9%)	266 (2.4%)	356 (7.9%)

THE COHORT STUDY complied with the principles established by the Declaration of Helsinki, and written informed consent was obtained by each subject. The ethical committee of the University of Milan (n. 6/2019) approved the study procedures.

Variables and measurements

OUTCOMES for supervised learning were both continuous and categorical. For categorical outcomes, important nutritional outcomes were chosen, either is subpopulation at risk or common to the whole sample; while dichotomizing such outcomes is challenging and potentially misleading, the main concern was to provide macro prediction on future outcome based on general practice employed in our specific setting, in the context of a CDSS.

- *continuous outcomes:* it was deemed useful to try to predict continuous outcomes generated through reference technique that, for specific reasons, cannot be used with every patient; that would permit to not rely in those cases on externally developed predictive equation, but to maximise accuracy through setting-specific algorithms:
  - resting energy expenditure (REE) by indirect calorimetry (kcal/day), the use of indirect calorimetry could be prevented
  - total body water (TBW) by bio-impedance analysis (l)
- *categorical outcomes:*
  - weight loss failure (0 = weight loss, 1 = no weight loss): weight loss was defined as reaching -5% of baseline body weight at follow-up
  - failure to improve basal glycemia (0 = improved, 1 = not improved): improvement was defined as recording fasting glucose >110 mg/dL at baseline and <100 mg/dL at follow-up
  - failure to improve total cholesterolemia (0 = improved, 1 = not improved): improvement was defined as recording total cholesterol >220 mg/dL at baseline and <200 mg/dL at follow-up
  - failure to improve triglyceridemia (0 = improved, 1 = not improved): improvement was defined as recording triglycerides >180 mg/dL at baseline and <150 mg/dL at follow-up

For prospective outcomes follow-up was defined as a period between 1 month and 7 months after baseline evaluation.

PREDICTORS were derived by all baseline measurements. These included:

- *demographic data:* age, sex, education, occupation, marital status
- *medical history:* family status, menstruation, pregnancies, diet status, diet history, physical activity, smoking, pharmacological treatments, clinical signs, weight history
- *parameters:* physical exam, blood pressure, anthropometry, bioimpedance analysis, ultrasound (abdomen fat thicknesses), indirect calorimetry, laboratory exams
- *questionnaires:* anxiety, depression, binge eating, emotion regulation, eating disorders and adherence to a Mediterranean diet

Overall the included predictors gave a variable representation of the following domains: biology, individual psychology, individual activity, activity environment, societal influences.

A detailed list of data collection procedures, variables included and their coding is available in the Appendix.

Machine learning and statistical analysis

For supervised learning maximum predictive strength was sought through optimization of relevant metrics. For continuous variables, the root-mean-square error (RMSE) and the coefficient of determination R^2; for categorical variables, the correct classification fraction (CCF) and the receiver operating characteristic area under the curve (AUROC). Between accuracy and discrimination ability, accuracy was selected as the most relevant metric in the clinical settings, and was sought through minimization of RMSE for continuous variables and maximization of the CCF for categorical variables.

Several statistical and machine learning models were compared using V-fold cross-validation resampling. For models requiring tuning parameters, a grid made of several combinations of tuning parameters was tested via V-fold cross-validation.

Prior to model selection, per-model preprocessing steps were defined in order to guarantee the best predictive ability for the specific model. To capture uncertainty about non-deterministic data manipulation, all preprocessing steps were repeated in each cross-validation fold.

Unsupervised learning was employed as an optional preprocessing step aimed to reduce the dimensionality of the dataset. In particular, principal component analysis (PCA) was used to transform the set of predictors in a reduced number of predictors designed to capture the maximum amount of information in the original variables. A potential benefit of this approach, other than the dimensionality reduction, is the production of statistically independent predictors that can ameliorate the problem of inter-variables correlations in the dataset.

Software used to perform the analysis was R 4.1.1 (R Core Team 2021). Refer to the Appendix for a detailed explanation of machine learning and statistical analysis used, and R code used to carry out the analysis.

Results

Model screening results for continuous and categorical outcomes, respectively, are shown in Figure 1 and Figure 2. For each outcome, a metric of accuracy and discrimination ability is presented (for continuous outcomes the RMSE and R^2, while for categorical outcomes the CCF and the AUROC). For each model and metric, the mean and confidence intervals obtained by the cross-validation process are shown (although, for each outcome and model, only the best combination of hyperparameters is shown). For each model, a version with and without PCA preprocessing (unsupervised learning) is shown.

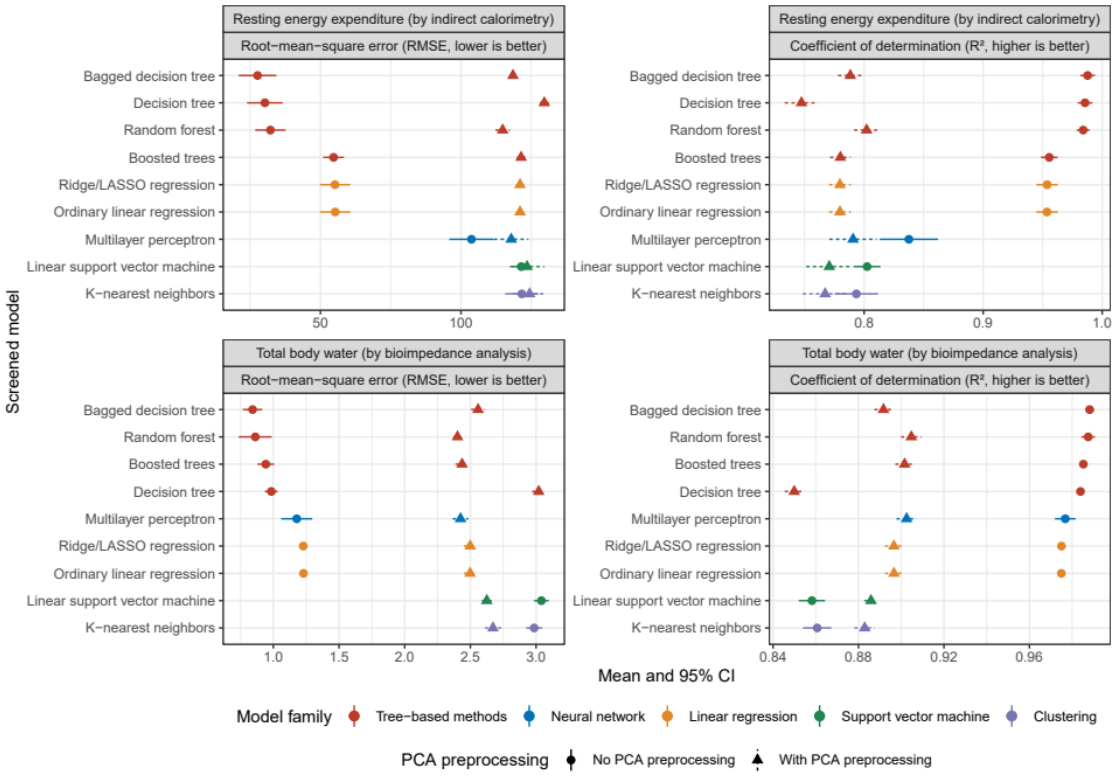

Figure 1: Comparison of accuracy (root-mean-square error, RMSE) and discrimination ability (coefficient of determination, R^2) of statistical and machine learning models in the prediction of continuous outcomes. For each model and metric, mean and confidence bounds (95% confidence) across resamples were computed. For each model, an alternative with and without principal component analysis (PCA, unsupervised learning) is shown.

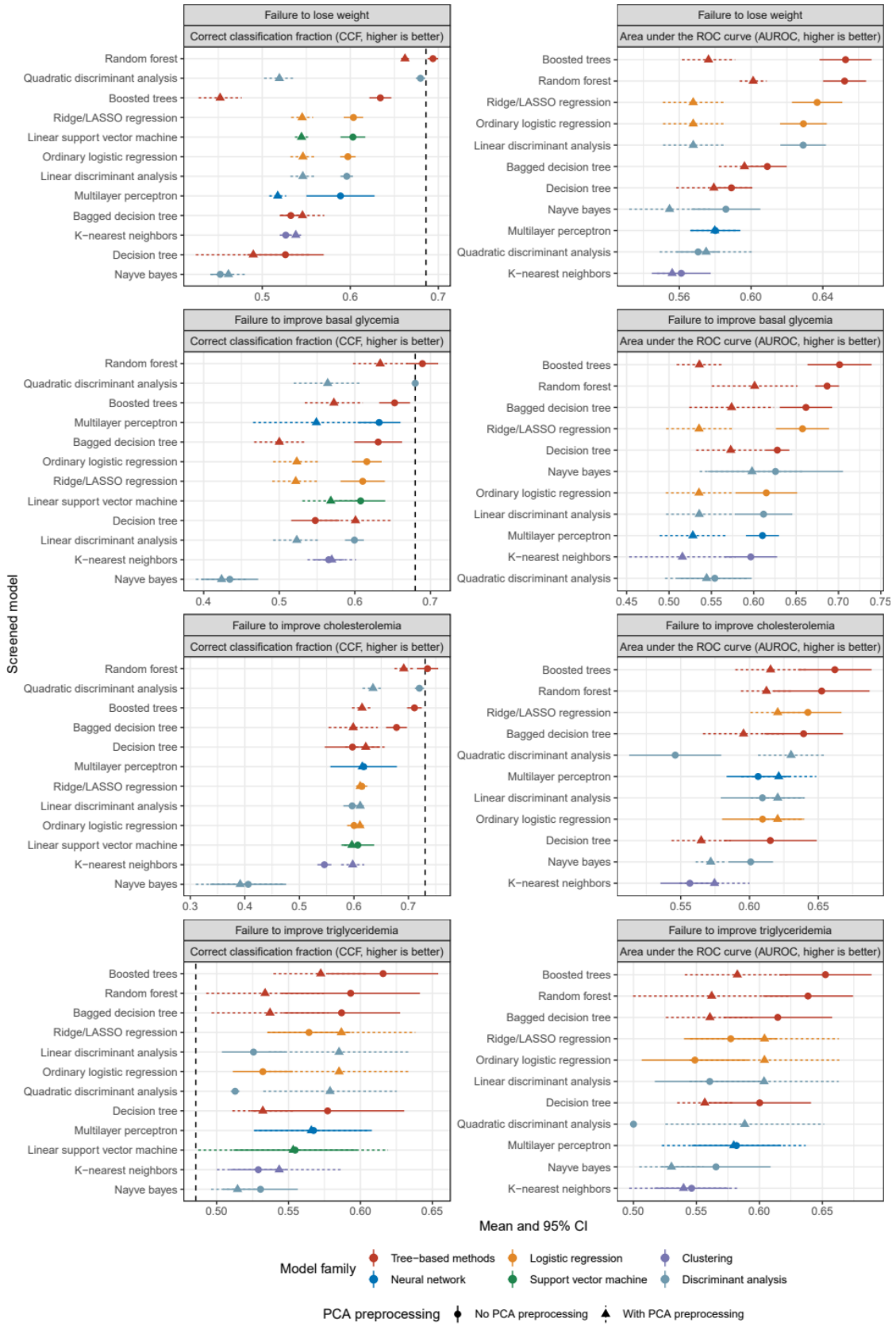

Figure 2: Comparison of accuracy (correct classification fraction, CCF) and discrimination ability (area under the ROC curve, ROC AUC) of statistical and machine learning models in the prediction of categorical outcomes. For each model and metric, mean and confidence bounds (95% confidence) across resamples were computed. For each model, an alternative with and without principal component analysis (PCA, unsupervised learning) is shown. For each outcome, the historical rate of patients experiencing the event is marked with a vertical dashed line.

For continuous outcomes, machine learning models based on decision trees (simple decision trees, bagged decision trees, boosted trees, and random forest) were generally the best performing models, producing both models with very low RMSE and a high R^2. PCA was generally not useful in improving the predictive ability of these models. Table 2 show in details accuracy and discrimination ability metrics for the best performing model (bagged decision trees for both outcomes) also including the hyperparameters that were selected in the screening process. Figure 3 compares measured outcomes with outcomes predicted from the best performing model, in a calibration plot obtained from the cross-validation process (the red dashed line represents the line of equality, where ideally the points would lie).

Table 2: Best model for each continuous outcome, ranked by accuracy (root-mean-square error, RMSE).

Outcome	Best model	Model hyperparameters	RMSE [1]	R^2 [1]
Resting energy expenditure (kcal)	Bagged decision tree	Cost/complexity parameter = 6.36e-07; Maximum depth = 15; Minimal node size = 34	27.6 (20.9, 34.3)	0.988 (0.982, 0.994)
Total body water (l)	Bagged decision tree	Cost/complexity parameter = 3.24e-10; Maximum depth = 9; Minimal node size = 8	0.842 (0.768, 0.916)	0.988 (0.986, 0.99)

[1] Mean (95% CI)

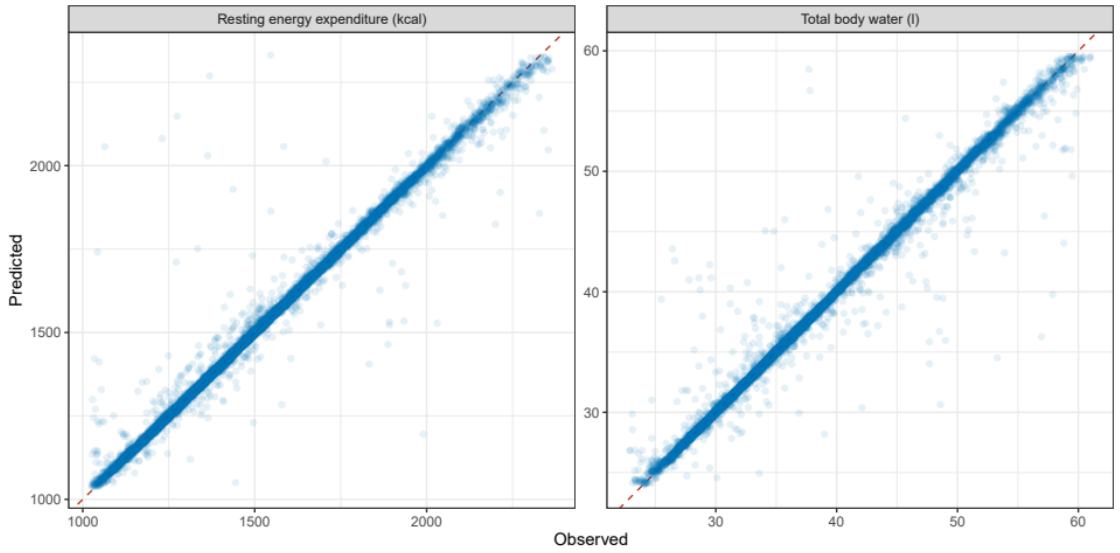

Figure 3: Calibration plot for continuous outcomes. The line of equality is drawn as a red dashed line.

For categorical outcomes, machine learning models based on decision trees (simple decision trees, bagged decision trees, boosted trees, and random forest) were generally the best performing models, producing both models with relatively high CCF and AUROC. PCA was generally not useful in improving the predictive ability of these models. The accuracy results need to be put in context of a naive classifier that would assume that all patients experience the event, which should have an accuracy similar to the historical proportion of events in the dataset (denoted in Figure 2 for each outcome with a vertical dashed line). Taking the naive classifier in consideration, only the best model of the weight outcome was able to consistently record a better performance of the naive classifier, while models predicting a (lack of) triglyceridemia improvement were generally able to perform better than the naive classifier. Table 3 show in details accuracy and discrimination ability metrics for the best performing model (random

forest for the cholesterolemia model, boosted trees for all other outcomes). Table 4 reports confusion matrices (true and false positive and negative rates) obtained from cross-validation of the best models for each outcome. Figure 4 shows the ROC curves for the best performing model for each outcome, showing the relationship between sensitivity and specificity over a continuum of different event thresholds (the dashed line denotes the expected performance of random guessing in an unbalanced setting).

Table 3: Best model for each categorical outcome, ranked by accuracy (correct classification fraction, CCF).

Outcome	**Best model**	**Model hyperparameters**	**Accuracy** [1]	**AUROC** [1]
Failure to lose weight	Random forest	# randomly selected predictors = 509; # trees = 1440; Minimal node size = 12	0.694 (0.688, 0.7)	0.652 (0.64, 0.664)
Failure to improve basal glycemia	Random forest	# randomly selected predictors = 379; # trees = 403; Minimal node size = 34	0.689 (0.669, 0.71)	0.687 (0.673, 0.701)
Failure to improve cholesterolemia	Random forest	# randomly selected predictors = 126; # trees = 1240; Minimal node size = 8	0.735 (0.715, 0.754)	0.653 (0.618, 0.687)
Failure to improve triglyceridemia	Boosted trees	# randomly selected predictors = 94; # trees = 363; Minimal node size = 27; Maximum depth of a tree = 5; Learning rate = 0.00171; Minimum loss reduction = 0.091; Proportion of observations sampled = 0.633; # iteration before stopping = 3	0.616 (0.577, 0.654)	0.652 (0.616, 0.689)

[1] Mean (95% CI)

Table 4: Confusion matrices for the best model of each categorical outcome.

	Failure to lose weight			**Failure to improve basal glycemia**			**Failure to improve cholesterolemia**			**Failure to improve triglyceridemia**		
	Predicted event	*Predicted no event*	**Total**	*Predicted event*	*Predicted no event*	**Total**	*Predicted event*	*Predicted no event*	**Total**	*Predicted event*	*Predicted no event*	**Total**
Event	5,172 (94%)	313 (5.7%)	5,485 (100%)	441 (88%)	58 (12%)	499 (100%)	1,439 (97%)	52 (3.5%)	1,491 (100%)	211 (70%)	91 (30%)	302 (100%)
No event	2,128 (85%)	383 (15%)	2,511 (100%)	167 (71%)	68 (29%)	235 (100%)	483 (88%)	67 (12%)	550 (100%)	148 (46%)	172 (54%)	320 (100%)
Total	7,300 (91%)	696 (8.7%)	7,996 (100%)	608 (83%)	126 (17%)	734 (100%)	1,922 (94%)	119 (5.8%)	2,041 (100%)	359 (58%)	263 (42%)	622 (100%)

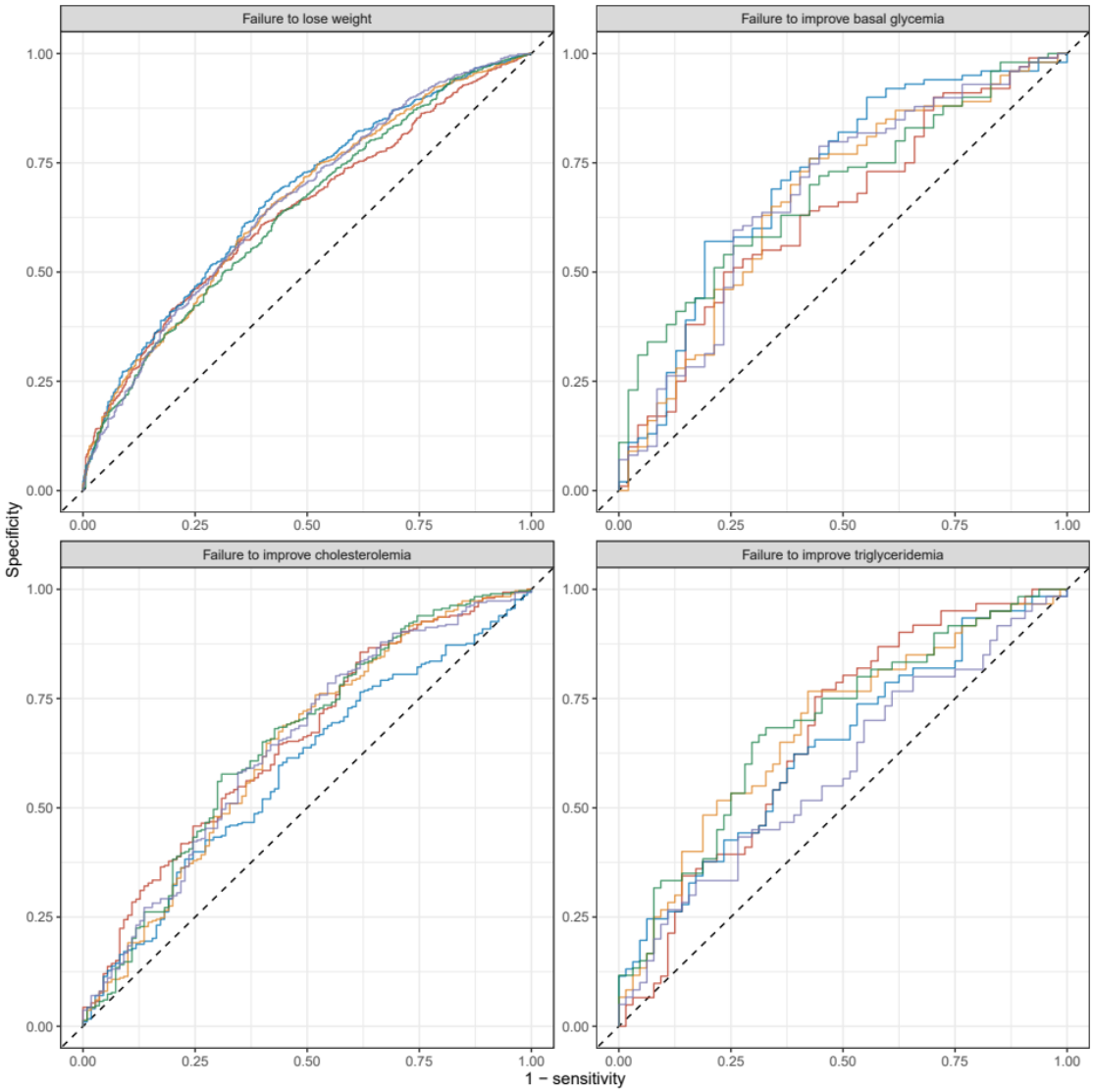

Figure 4: Receiver operator curve for categorical outcomes. Each resample is drawn with a different color.

Discussion

Our results show good overall performance of machine learning models when applied to clinical data in nutritional settings. Supervised learning algorithms performed extremely well in the prediction of continuous cross-sectional outcomes, showing great potential as compensatory system for missing variables that could aid the clinician with a better prediction than those provide by externally developed equations or surrogate methods, when a particular patient does not satisfy the eligibility criteria for the reference method. In the prediction of prospective categorical outcomes, machine learning models were not only the best performers, they were the only one able to provide better predictions than a naive classifier based on historical rates in highly unbalanced settings. As such, they show potential as an alert system aimed to improve identification of non-responders based on baseline evaluation and setting-specific historical success rates, but to date their discrimantion ability remains low. Unsupervised learning, in the form of dimensionality reduction preprocessing, was

generally not useful in improving predictive performance of supervised machine learning models.

In the context of a CDSS, prediction of cross-sectional variables, such as those selected in this thesis (REE and TBW), can provide useful information to the clinician when circumstances arise that prevent the use of the reference technique for some patients. Indirect calorimetry is considered the gold standard method for measuring REE and can be routinely used in clinical practice, but there are instances in normal clinical practice in which it cannot be used (eg. claustrophobic patients, lack of fasting, interfering pharmacological treatments). The same applies to bioimpedance analysis, a technique that can provide useful information on hydration and body composition through the measurement of TBW, but is not advised when the patient is carrying an implanted cardio-verter-defibrillator, pace makers, prostheses or metal implants. In cases in which these techniques cannot be used, the clinician usually relies on externally developed predictive equations. These equations have the advantage of having been developed with a rigorous and specific study design, but the need of being highly transferable from the original development environment to other foreign settings imposes several constraints. The equations usually only include a few important predictors of the outcome in order to be more broadly used, but doing so they cannot fully take advantage of highly-dimensional data that is nonetheless collected for the patient. Also, the population on which these equations have been developed do not necessarily match the population object of our clinical practice.

On the other hand, the selected classification tasks were certainly more ambitious, being prospective and considering outcomes easily influenced by several factors. The weight, glycemia and cholesterolemia models were also challenged by highly unbalanced outcomes that, even after upsampling in the preprocessing phase, produced models with good sensitivity but poor specificity. The most interesting result came from the weight model, as it was the only one in this unbalanced setting to have consistently beaten the naive classifier based on historical rates by a few percentage points. Prediction of weight loss is certainly one of the most challenging issues in nutrition, as after decades of research, it remains an untackled problem. In the prediction of the categorical outcomes, the weight loss model was the one with the largest sample size, being an outcome routinely explored and documented in our patients. This has likely contributed to the difference in performance versus the glycemia and cholesterolemia model. The model in the triglyceridemia prediction was the one that more consistently performed better than the naive classifier. Noticeably this was the most balanced setting and that seems to have contributed positively in this regard. Once again these models do not increase the data collection burden as they are based on routinely collected clinical data, and when integrated in a CDSS, they can leverage these data to provide the clinician with a probability of success on several relevant outcomes based on historical data of patients similar to the one at hand. These results may be included in a CDSS in the form of probability of the patient being a non-responder to historical treatment strategies adopted for similar patients. More than a binary classifier, showing the probability of a patient being a non-responder would allow the clinician to measure the confidence of the suggestion provided by the machine learning algorithm and integrate the result in her own clinical decision process.

INTERPRETABILITY AND UNCERTAINTY are one of most researched topics today in AI and machine learning. Some computational reasoning methods in AI, such as neural networks, are

considered black boxes to end-users. Auditing has been proposed as a pragmatic approach to evaluating opaque algorithms that were devised autonomously. This follows an analogy to human judgement; typically we measure outcomes, not problem-solving style or cognitive process. However, given the fundamental healthcare ethic of "first do no harm", some authors argue that more effort is needed in the design phases to explain the principles of a computational model to allow transparent assessment (Magrabi et al. 2019). They suggest this would help to keep clinicians and patients engaged and avoid conflict between practitioners and commercial algorithm developers. They advise algorithm developers, including those who operate on a proprietary basis, on the need to consider how to open the black box (even if partially) and work within a framework for shareable biomedical knowledge so that clinicians can judge the merits of AI models.

In this thesis, interpretability of the best performing models was limited by the complexity of the model structure. The best performing models were all tree-based models, that is models that split the data multiple times according to certain cutoff values in the predictors. Through splitting, different subsets of the dataset are created, with each instance belonging to one subset. The final subsets are called terminal or leaf nodes and the intermediate subsets are called internal nodes or split nodes. To predict the outcome in each leaf node, the average outcome of the training data in this node is used, for both categorical and continuous outcomes.

For basic decision tree models, the interpretation is simple: starting from the root node, the path to the next node and the edges tell which subsets is going to be selected, and when the leaf node is reached, the node tells the predicted outcome. Also, the overall importance of a predictor in a decision can be computed by going through all the splits for which the predictor was used and measuring how much it has reduced the variance or other indices compared to the parent node. The sum of all importances is scaled to 100 so that each importance can be interpreted as a share of the overall model importance. Individual predictions of a decision tree can be explained by decomposing the decision path into one component per predictor.

While simple decision trees have natural visualization and tend to create good explanations for the prediction produced, they have several drawbacks. Simple decision trees fail to deal with linear relationships, as any linear relationship between an input predictor and the outcome has to be approximated by splits, creating an inefficient step function. They lack smoothness and so slight changes in the input predictors can have a big impact on the predicted outcome, which is usually not desirable. Trees are also quite unstable, because each split depends on the parent split and so few changes in the training dataset can create a completely different tree. Also, decision trees are very interpretable as long as they are short, as the number of terminal nodes increases quickly with depth.

To overcome the disadvantage of simple decision trees and increasing their predictive accuracy and robustness, several techniques have been developed, such as bagging, random forests, and boosting. Bagging improves both accuracy and stability of decision trees by generating several bootstrapped training data sets on which to fit simple decision trees that are then averaged in a single model. While bagging does improve overall on simple decision trees, it can suffer from strong predictors in the dataset that tend to create very similar trees even in the bootstrapped data sets. Random forest improves on bagging by forcing each split to consider a subset of the predictors,

effectively forcing decorrelation between trees. Boosting is similar in concept to bagging, but instead of growing trees on different bootstraps resamples, it grows trees sequentially based on residuals of previously fitted trees. Sequential trees in boosting are forced to grow small and so the overall tree grows slowly and more stable than a single decision tree.

Our results confirm the benefits of these approaches to improve predictive accuracy and stability across resamples of decision trees, at the expense of the clear interpretability of simple decision trees or other models. While some innovative approaches have been developed to aid interpretability of complex tree-base models (Friedman and Popescu 2008), the high-dimensionality and more importantly the study design used in this thesis limit the usefulness of any tentative interpretation of the fitted models, potentially leading to misleading conclusions.

BUT IS INTERPRETABILITY NECESSARY? Several assumptions underlie the focus on explainability as the primary means by which to address concerns around accountability, transparency, trust, and adoption of machine learning in healthcare (Sendak et al. 2020):

- Clinicians are presumed to have substantial technical and quantitative expertise with which to engage with explainable machine learning.
- Clinicians often incorporate information into clinical decisions without a comprehensive understanding of the mechanism by which the information is generated.
- The volume and complexity of the knowledge that physicians need to master has grown exponentially beyond their capacity as individuals and as such the effective use of information in clinical decision making should be prioritised rather than comprehensive understanding of how information is generated.

The application of medical knowledge does not necessarily require the identification of causal relations. The human body is in many ways “a black box”, in which the causes and mechanisms of illnesses often elude explanation. The use of AI fundamentally calls into question the extent to which we tolerate uncertainty in medical decision making (Harish et al. 2021). Even with AI to help clinicians weigh the likelihood of various diagnoses (and the usefulness of various treatments) against one another, it is not possible to reduce diagnostic uncertainty to zero. Successful integration of AI into the clinical decision-making framework requires clinicians to handle uncertainty as a relative measure rather than an absolute value to minimize.

THIS THESIS offers a glimpse of the potential of machine learning algorithms applied to routinely collected clinical data. The strengths of this thesis lie in the dimension of the dataset used, both as number of observations and number of variables/domains explored. Also, many algorithms were tested both among machine learning and statistical algorithms, trying to optimize each one with a series of selected per-model pre-processing steps. On the other hand, some limitations have to be noted. First, while the dataset was certainly big by traditional standards, the number of observation relative to the number of predictors was not deemed infinite and to avoid selection bias derived from splitting the dataset in training-and-test splits, only internal cross-validation was used to test the predictive abilities of the models. Also, in the prediction of categorical prospective outcomes, a detailed encoding of all aspects of the

therapeutic plan was not available, but it would have likely improved the prediction and would have surely provided a much greater insight for the clinician having to choose between different strategies.

In conclusion, AI thinking processes do not mirror how a human processes questions. Humans have an immediate instinct for whether they know the correct answer, and this intuitive confidence is a subjective experience for a human. While both humans and AI take confidence-driven approaches, only AI explicitly incorporates confidence as a quantifiable and objective metric. This can lead to instances in which a (low-confidence) AI conclusion is obviously wrong from the human viewpoint, explaining why the public may be uncomfortable with an AI system functioning under uncertainty. For a system to wield decision-making power, one must accept that the AI system will eventually draw incorrect inferences and that humans using intuition will see these incorrect inferences as blatantly obvious. On the other hand, acting on high-confidence suggestions of machine learning algorithms and correctly encoding the new therapeutic approaches will eventually improve the predictive ability of these models that are dynamic and evolving in nature.

The clinical adoption of AI may be a reflection of how intrinsic uncertainty is to medicine. Clinicians must reckon with and ultimately accept the fact that no diagnosis is certain, which is why they synthesize differential diagnoses. The calculated probabilities of AI-based CDSSs must, in practice, be reconciled with the intuition of expert clinicians if we are to understand differences in how recommendations emerge.

Appendix

Data collection

Demographic data. Demographic data was self-reported by the patient and included age, sex, education, occupation, and marital status.

Medical history. An accurate medical interview was carried out, along with the collection of medical condition self-reported diagnosis and information regarding current drug therapies. For women, menstruation status and pregnancy history was investigated. Smoking habits, dietary habits, diet history and weight history were investigated. A structured interview was employed to investigate physical activity levels. Subjects engaging in any structured physical exercise for >2 h/week were deemed as physically active.

Anthropometry. Body weight, body height and fat mass fraction were assessed with anthropometric methods. All anthropometric measurements were collected by well-trained registered dietitians at the ICANS center. Procedures used are detailed in Lohman and Roche (1988).

Body weight was measured to the nearest 100 g with a mechanical column scale graduated to 100 g, and with a capacity of 160 kg (Seca 700, Seca GmbH, Hamburg, Germany). Body height was measured to the nearest 1 mm with a stadiometer graduated to 1 mm, and with a measuring range of 20-205 cm (Seca 217, Seca GmbH, Hamburg, Germany). Waist circumference was measured to the nearest 0.1 cm with an inextensible metric tape, wide 0.5 cm, and graduated to 1 mm (Gima 27341, Gima S.p.A., Gessate, Italy). Biceps, triceps, subscapular, and suprailiac skinfold thicknesses were measured to the nearest 0.1 mm using a skinfold caliper with a 35 mm^2 jaw face area, exerting a 10±2 g/mm^2 pressure between the jaws, with a range of 0-40 mm, calibrated to 0.2 mm (Holtain Tanner/Whitehouse Skinfold Caliper, Croswell, UK).

Body mass index (BMI, kg/m^2) was calculated as body weight divided by the square of the body height, and classified according to the World Health Organization (2000) guidelines. Body density (kg/l) and fat mass fraction (as %) were estimated using formula provided by Durnin and Womersley (1974) and Siri (1961), respectively.

Bioimpedance analysis. Impedance was measured using a tetrapolar 8-point tactile electrode system (InBody 720, Biospace, Seoul, Korea) at 1, 5, 50, 250, 500 and 1000 kHz. The system measured the impedance of the participant's right arm, left arm, trunk, right leg and left leg. Participants stood on the scale platform of the instrument and grasped the handles of the device, to provide contact with a total of eight electrodes (two for each foot and for each hand).

Manufacturer equations were used to estimate the following variables: total body water (TBW), extracellular water (ECW) and intracellular water (ICW).

The intra-examination coefficient of variation for bio-impedance analysis was 0.8 %.

Ultrasound. Visceral and subcutaneous adipose tissue thicknesses were measured at the abdominal level by the same operator following a validated standardized protocol.

The instrument used was a Logiq 3 Pro system for abdominal ultrasonography, equipped with a 3.5 MHz convex-array probe and a 7.5 MHz linear probe (GE Healthcare, Milwaukee, WI, USA). Visceral adipose tissue was measured as the distance between the posterior surface of the rectus abdominis muscle and the anterior wall of the aorta at the level of linea alba, and subcutaneous adipose tissue was measured as the distance between the external face of the rectus abdominis muscle and the epidermis. Both thicknesses were determined three times, one centimetre above the umbilicus, and a mean measurement was computed. The intra-operator coefficient of variation for repeated VAT and SAT measurements in our laboratory is 0.8%.

Indirect calorimetry. Resting energy expenditure was measured at baseline with indirect calorimetry. All measurements were performed early in the morning, after a 12-hours fast, with subjects lying supine, at rest but awake, in a quiet and thermally neutral environment (24 °C). After a 15 minutes resting period, O_2 consumption and CO_2 production were measured using the canopy dilution technique (Ferrannini 1988), with patients wearing a transparent ventilated canopy for 30 minutes, sampling gases every 30 seconds. To avoid gas leakages, the subject's head was carefully wrapped with a veil.

Technical details of the indirect calorimeter used (Q-NRG+, Cosmed srl, Rome, Italy) are presented by Delsoglio et al. (2020). Calibration of the flowmeter and gas analyzers were performed according to the manufacturer's instructions and schedule.

For each measurement, data collected during the first 5 min were discarded, while data collected during the remaining minutes were averaged (mean) and scaled to provide daily resting gas exchanges. Steady state was defined as the first five consecutive stable 30 seconds readings with a coefficient of variation <10 % for VO_2 and VCO_2, and when available, data in steady state were preferred. The Weir formula (V. Weir 1949) was used to estimate resting energy expenditure from gas exchanges measured at rest by indirect calorimetry.

Laboratory exams. Fasting blood samples were taken by venipuncture of the antecubital vein using vacuum tubes, in either sitting or lying position. After centrifugation (800g × 10 min at 5°), aliquots of samples were stored at -80° until further analysis. Urine samples of the second urination of the day were collected measuring time from first and second urination with a timer. Pre-prandial glycemia and ketonemia were self-measured by each patient with an in vitro diagnostic medical device for blood glucose and β-ketone self-testing (GlucoMen LX PLUS, Menarini Diagnostics).

An auto-analyzer (Cobas Integra 400 plus, Roche Diagnostics, Mannheim) was used to determine serum glucose, HBA1C, cholesterol, high-density lipoprotein (HDL) cholesterol, low-density lipoprotein (LDL) cholesterol, triglycerides, glutamic oxaloacetic transaminase (GOT), glutamate-pyruvate transaminase (GPT), gamma-glutamyltransferase (GGT) and urinary urea.

Questionnaires. Anxiety, depression, binge eating, emotion regulation, eating disorders and adherence to a Mediterranean diet were assessed at baseline with the State-Trait Anxiety Inventory (Spielberger 2010), Depression Questionnaire (Vidotto et al. 2010), Binge Eating Scale (Gormally et al. 1982), Difficulties in Emotion Regulation Scale (Dan-Glauser and Scherer 2012), Eating Disorder Inventory (Garner, Olmstead, and

Polivy 1983), and the Mediterranean Diet Adherence Screener (Martínez-González 2012).

All questionnaires were self-administered by the patient.

We used the Form X of the State-Trait Anxiety Inventory, the original version of BES, and version 3 of the Eating Disorder Inventory.

Continuous and categorical scores were computed following the scoring instructions of each questionnaire.

Preprocessing

THE PREPROCESSING STEPS were tailored to the type of model being fit. Overall the following preprocessing steps were used:

- *dummy*: transforming qualitative predictors with a numeric encoding
- *zv*: removing predictors with low variance
- *impute*: estimating missing variables (in model selection phase only simple imputation was performed to limit computational time: for continuous variable median imputation, for categorical variables a new "unknown" level
- *decorrelate*: filtering out highly-correlated predictors or using principal component analysis to reduce dimensionality, or using a model-based technique (e.g. regularization)
- *normalize*: centering and scaling of predictors
- *transform*: transforming predictors to be more symmetric

The table below shows preprocessing steps performed for each model.

Model selection

FOR SUPERVISED LEARNING, the following models were evaluated for both type of outcome (unless stated otherwise):

- *linear regression* (continuous outcomes only)
- *logistic regression* (categorical outcomes only)
- *linear discriminant analysis* (categorical outcomes only)
- *quadratic discriminant analysis* (categorical outcomes only)
- *naive Bayes* (categorical outcomes only), tuned for kernel smoothness, and Laplace correction
- *K-nearest neighbour*, tuned for number of nearest neighbors, and distance weighting function, Minkowski distance order
- *ridge regression and LASSO*, tuned for the amount of regularization, and the proportion of LASSO penalty
- *decision trees*, tuned for tree depth, minimal node size, and cost-complexity parameter
- *bagged trees*, tuned for the cost/complexity parameter used by CART models, the maximum depth of a tree, the minimum number of data points in a node that are

required for the node to be split further, and a cost value to assign to the class corresponding to the first factor level

- *random forest*, tuned for number randomly selected predictors, number of trees, and minimal node size
- *boosted trees*, tuned for tree depth, the number trees, the learning rate, the number randomly selected predictors, the minimal node size,the minimum loss reduction,the proportion observations sampled, and the number iterations before stopping
- *linear support vector machine*, tuned for cost, and insensitivity margin
- *single layer neural network*, tuned for the number of hidden units, the amount of regularization, and the number of epochs

To limit computation time, 5-fold cross validation and 10 values per hyperparameter were tested. Where possible, racing with ANOVA models (Kuhn 2014) was performed to further reduce computation time.

Model preprocessing, turning, resampling, and fitting were performed with the notable addition of the `Tidymodels` package (Kuhn and Wickham 2020) to R (R Core Team 2021).

Code for preprocessing procedures

```
.recipe <-
  list()

.recipe$preprocessing <-
  function(recipe) {
    recipe %>%
      step_date(collected_on, features = c("month")) %>%
      step_rm(patient_id, where(lubridate::is.Date),
              where(lubridate::is.timepoint), contains("value")) %>%
      step_mutate_at(where(is.logical), fn = ~as.factor(.x) %>%
                       forcats::fct_relabel(janitor::make_clean_names)) %>%
      step_mutate_at(where(is.numeric),
                     fn = ~ ifelse(
                       .x < quantile(.x, probs = .01, na.rm = T) |
                         .x > quantile(.x, probs = .99, na.rm = T),
                       NA_real_,
                       .x
                     )) %>%
      step_mutate_at(where(is.numeric) & !starts_with("ders"),
                     fn = ~ ifelse(.x < 0, NA_real_, .x))
  }

.recipe$imputing <-
  function(recipe) {
    recipe %>%
      step_impute_median(all_numeric_predictors()) %>%
      step_novel(all_nominal_predictors()) %>%
      step_unknown(all_nominal_predictors()) %>%
      step_naomit(all_outcomes(), skip = T)
  }

.recipe$dummy.nzv.impute.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
```

```
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.upsample.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.upsample.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$nzv.impute.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$nzv.impute.upsample.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$nzv.impute.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
```

```
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$nzv.impute.upsample.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$nzv.upsample.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      step_naomit(all_outcomes(), skip = T) %>%
      step_upsample(all_outcomes()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$nzv.decorrelate <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      step_naomit(all_outcomes(), skip = T) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$nzv.impute.upsample.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$nzv.impute.decorrelate.normalize.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_BoxCox(all_numeric_predictors() & !starts_with("ders")) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
```

```
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.decorrelate.normalize.transform <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_BoxCox(all_numeric_predictors() & !starts_with("ders")) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_upsample(all_outcomes()) %>%
      step_BoxCox(all_numeric_predictors() & !starts_with("ders")) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }

.recipe$dummy.nzv.impute.decorrelate.normalize.transform.pca <-
  function(recipe) {
    recipe %>%
      .recipe$preprocessing() %>%
      .recipe$imputing() %>%
      step_BoxCox(all_numeric_predictors() & !starts_with("ders")) %>%
      step_normalize(all_numeric_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_other(all_nominal_predictors(), other = "low_freq_values") %>%
      step_dummy(all_nominal_predictors()) %>%
      step_nzv(all_predictors()) %>%
      step_corr(all_numeric_predictors()) %>%
      step_pca(all_numeric_predictors())
  }
```

Code for linear regression tasks

```
regr_wflow <-
  function(sample) {
    list(
      race = bind_rows(
        workflow_set(
          list(
            dummy.nzv.impute.decorrelate =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.decorrelate(),
            dummy.nzv.impute.decorrelate.normalize.pca =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.decorrelate.normalize.pca()
          ),
          list(
            linearreg.lm =
```

```
      linear_reg() %>%
      set_engine("lm"),
    linearreg.glmnet =
      linear_reg(penalty = tune(),
                 mixture = tune()) %>%
      set_engine("glmnet"),
    trees.boost =
      boost_tree(
        tree_depth = tune(),
        trees = tune(),
        learn_rate = tune(),
        mtry = tune(),
        min_n = tune(),
        loss_reduction = tune(),
        sample_size = tune(),
        stop_iter = tune()
      ) %>%
      set_engine("xgboost") %>%
      set_mode("regression")
  )
),
workflow_set(
  list(
    dummy.nzv.impute.decorrelate.normalize.transform =
      recipe(outcome ~ ., sample) %>%
      .recipe$dummy.nzv.impute.decorrelate.normalize.transform(),
    dummy.nzv.impute.decorrelate.normalize.transform.pca =
      recipe(outcome ~ ., sample) %>%
      .recipe$dummy.nzv.impute.decorrelate.normalize.transform.pca()
  ),
  list(
    nearestneighbor =
      nearest_neighbor(
        neighbors = tune(),
        weight_func = tune(),
        dist_power = tune()
      ) %>%
      set_engine("kknn") %>%
      set_mode("regression"),
    nnet.mlp =
      mlp(
        hidden_units = tune(),
        penalty = tune(),
        epochs = tune()
      ) %>%
      set_engine("nnet") %>%
      set_mode("regression"),
    svm.linear =
      svm_linear(cost = tune(), margin = tune()) %>%
      set_engine("LiblineaR") %>%
      set_mode("regression")
  )
),
workflow_set(
  list(
    nzv.decorrelate =
      recipe(outcome ~ ., sample) %>%
      .recipe$nzv.decorrelate(),
    nzv.impute.decorrelate.normalize.pca =
      recipe(outcome ~ ., sample) %>%
      .recipe$nzv.impute.decorrelate.normalize.pca()
  ),
  list(
    trees.decision =
      decision_tree(
        tree_depth = tune(),
        min_n = tune(),
        cost_complexity = tune()
      ) %>%
      set_engine("rpart") %>%
      set_mode("regression"),
```

```
          trees.bag =
            bag_tree(
              cost_complexity = tune(),
              tree_depth = tune(),
              min_n = tune()
            ) %>%
            set_engine("rpart") %>%
            set_mode("regression")
        )
      )
    ),
    no_race =
      workflow_set(
        list(
          nzv.impute.decorrelate =
            recipe(outcome ~ ., sample) %>%
            .recipe$nzv.impute.decorrelate(),
          nzv.impute.decorrelate.normalize.pca =
            recipe(outcome ~ ., sample) %>%
            .recipe$nzv.impute.decorrelate.normalize.pca()
        ),
        list(
          trees.randforest =
            rand_forest(
              mtry = tune(),
              trees = tune(),
              min_n = tune()
            ) %>%
            set_engine("ranger") %>%
            set_mode("regression")
        )
      )
  )
}

regr_resmpl <-
  function(sample) {
    list(
      regr_wflow(sample)$race %>%
        workflow_map(
          "tune_race_anova",
          verbose = T,
          resamples = vfold_cv(sample, v = 5),
          grid = 10,
          control = control_race(
            verbose = T,
            save_pred = T,
            save_workflow = T
          )
        ),
      regr_wflow(sample)$no_race %>%
        workflow_map(
          verbose = T,
          resamples = vfold_cv(sample, v = 5),
          grid = 10,
          control = control_grid(
            verbose = T,
            save_pred = T,
            save_workflow = T
          )
        )
    )
  }

regr_resampled <-
  samples$reg %>%
  map(~regr_resmpl(.x))
```

Code for classification tasks

```
class_wflow <-
  function(sample) {
    list(
      race = bind_rows(
        workflow_set(
          list(
            dummy.nzv.impute.upsample.decorrelate =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.upsample.decorrelate(),
            dummy.nzv.impute.upsample.decorrelate.normalize.pca =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.upsample.decorrelate.normalize.pca()
          ),
          list(
            logisticreg.glm =
              logistic_reg() %>%
              set_engine("glm"),
            logisticreg.glmnet =
              logistic_reg(penalty = tune(),
                           mixture = tune()) %>%
              set_engine("glmnet"),
            discrim.linear =
              discrim_linear() %>%
              set_engine("MASS"),
            discrim.quad =
              discrim_quad() %>%
              set_engine("MASS"),
            trees.boost =
              boost_tree(
                tree_depth = tune(),
                trees = tune(),
                learn_rate = tune(),
                mtry = tune(),
                min_n = tune(),
                loss_reduction = tune(),
                sample_size = tune(),
                stop_iter = tune()
              ) %>%
              set_engine("xgboost") %>%
              set_mode("classification")
          )
        ),
        workflow_set(
          list(
            nzv.impute.upsample.decorrelate =
              recipe(outcome ~ ., sample) %>%
              .recipe$nzv.impute.upsample.decorrelate(),
            nzv.impute.upsample.decorrelate.normalize.pca =
              recipe(outcome ~ ., sample) %>%
              .recipe$nzv.impute.upsample.decorrelate.normalize.pca()
          ),
          list(
            discrim.naivebayes =
              naive_Bayes(smoothness = tune(),
                          Laplace = tune()) %>%
              set_engine("klaR")
          )
        ),
        workflow_set(
          list(
            dummy.nzv.impute.upsample.decorrelate.normalize.transform =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform(),
            dummy.nzv.impute.upsample.decorrelate.normalize.transform.pca =
              recipe(outcome ~ ., sample) %>%
              .recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform.pca()
          ),
```

```
      list(
        nearestneighbor =
          nearest_neighbor(
            neighbors = tune(),
            weight_func = tune(),
            dist_power = tune()
          ) %>%
          set_engine("kknn") %>%
          set_mode("classification"),
        nnet.mlp =
          mlp(
            hidden_units = tune(),
            penalty = tune(),
            epochs = tune()
          ) %>%
          set_engine("nnet") %>%
          set_mode("classification")
      )
    ),
    workflow_set(
      list(
        nzv.upsample.decorrelate =
          recipe(outcome ~ ., sample) %>%
          .recipe$nzv.upsample.decorrelate(),
        nzv.impute.upsample.decorrelate.normalize.pca =
          recipe(outcome ~ ., sample) %>%
          .recipe$nzv.impute.upsample.decorrelate.normalize.pca()
      ),
      list(
        trees.decision =
          decision_tree(
            tree_depth = tune(),
            min_n = tune(),
            cost_complexity = tune()
          ) %>%
          set_engine("rpart") %>%
          set_mode("classification"),
        trees.bag =
          bag_tree(
            cost_complexity = tune(),
            tree_depth = tune(),
            min_n = tune(),
            class_cost = tune()
          ) %>%
          set_engine("rpart") %>%
          set_mode("classification")
      )
    )
  ),
  no_race =
    workflow_set(
      list(
        nzv.impute.upsample.decorrelate =
          recipe(outcome ~ ., sample) %>%
          .recipe$nzv.impute.upsample.decorrelate(),
        nzv.impute.upsample.decorrelate.normalize.pca =
          recipe(outcome ~ ., sample) %>%
          .recipe$nzv.impute.upsample.decorrelate.normalize.pca()
      ),
      list(
        trees.randforest =
          rand_forest(
            mtry = tune(),
            trees = tune(),
            min_n = tune()
          ) %>%
          set_engine("ranger") %>%
          set_mode("classification")
      )
    ),
  accuracy_only =
```

```
      workflow_set(
        list(
          dummy.nzv.impute.upsample.decorrelate.normalize.transform =
            recipe(outcome ~ ., sample) %>%
            .recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform(),
          dummy.nzv.impute.upsample.decorrelate.normalize.transform.pca =
            recipe(outcome ~ ., sample) %>%
            .recipe$dummy.nzv.impute.upsample.decorrelate.normalize.transform.pca()
        ),
        list(
          svm.linear =
            svm_linear(cost = tune()) %>%
            set_engine("LiblineaR") %>%
            set_mode("classification")
        )
      )
  )
}

class_resmpl <-
  function(sample) {
    list(
      class_wflow(sample)$race %>%
        workflow_map(
          "tune_race_anova",
          verbose = T,
          resamples = vfold_cv(sample, v = 5, strata = "outcome"),
          grid = 10,
          control = control_race(
            verbose = T,
            save_pred = T,
            save_workflow = T
          )
        ),
      class_wflow(sample)$no_race %>%
        workflow_map(
          verbose = T,
          resamples = vfold_cv(sample, v = 5, strata = "outcome"),
          grid = 10,
          control = control_grid(
            verbose = T,
            save_pred = T,
            save_workflow = T
          )
        ),
      class_wflow(sample)$accuracy_only %>%
        workflow_map(
          "tune_race_anova",
          verbose = T,
          metrics = metric_set(accuracy),
          resamples = vfold_cv(sample, v = 5, strata = "outcome"),
          grid = 10,
          control = control_race(
            verbose = T,
            save_pred = T,
            save_workflow = T
          )
        )
    )
  }

class_resampled <-
  samples$class %>%
  map(~class_resmpl(.x))
```

References

Breiman, Leo. 2001. “Statistical Modeling: The Two Cultures (with Comments and a Rejoinder by the Author).” *Statistical Science* 16 (3): 199–231. https://doi.org/10.1214/ss/1009213726.

Chawla, Nitesh V., and Darcy A. Davis. 2013. “Bringing Big Data to Personalized Healthcare: A Patient-Centered Framework.” *Journal of General Internal Medicine* 28 (3): 660–65. https://doi.org/10.1007/s11606-013-2455-8.

Dan-Glauser, Elise S, and Klaus R Scherer. 2012. “The Difficulties in Emotion Regulation Scale (DERS).” *Swiss Journal of Psychology*.

Delsoglio, Marta, Yves Marc Dupertuis, Taku Oshima, Mart van der Plas, and Claude Pichard. 2020. “Evaluation of the Accuracy and Precision of a New Generation Indirect Calorimeter in Canopy Dilution Mode.” *Clinical Nutrition* 39 (6): 1927–34.

Deo, Rahul C. 2015. “Machine Learning in Medicine.” *Circulation* 132 (20): 1920–30. https://doi.org/10.1161/circulationaha.115.001593.

Docherty, Annemarie B., and Nazir I. Lone. 2015. “Exploiting Big Data for Critical Care Research.” *Current Opinion in Critical Care* 21 (5).

Durnin, J. V. G. A., and J. Womersley. 1974. “Body Fat Assessed from Total Body Density and Its Estimation from Skinfold Thickness: Measurements on 481 Men and Women Aged from 16 to 72 Years.” *British Journal of Nutrition* 32 (01): 77–97.

Ferrannini, Eleuterio. 1988. “The Theoretical Bases of Indirect Calorimetry: A Review.” *Metabolism* 37 (3): 287–301.

Friedman, Jerome H, and Bogdan E Popescu. 2008. “Predictive Learning via Rule Ensembles.” *The Annals of Applied Statistics* 2 (3): 916–54.

Frood, Sarah, Lee M. Johnston, Carrie L. Matteson, and Diane T. Finegood. 2013. “Obesity, Complexity, and the Role of the Health System.” *Current Obesity Reports* 2 (4): 320–26. https://doi.org/10.1007/s13679-013-0072-9.

Garner, David M, Marion P Olmstead, and Janet Polivy. 1983. “Development and Validation of a Multidimensional Eating Disorder Inventory for Anorexia Nervosa and Bulimia.” *International Journal of Eating Disorders* 2 (2): 15–34.

Gormally, Jim, Sionag Black, Sandy Daston, and David Rardin. 1982. “The Assessment of Binge Eating Severity Among Obese Persons.” *Addictive Behaviors* 7 (1): 47–55.

Harish, Vinyas, Felipe Morgado, Ariel D. Stern, and Sunit Das. 2021. “Artificial Intelligence and Clinical Decision Making: The New Nature of Medical Uncertainty.” *Academic Medicine* 96 (1).

Institute of Medicine. 2011. *Digital Infrastructure for the Learning Health System: The Foundation for Continuous Improvement in Health and Health Care: Workshop Series*

Summary. Edited by Claudia Grossmann, Brian Powers, and J. Michael McGinnis. Washington, DC: The National Academies Press. https://doi.org/10.17226/12912.

Jafari, Mohieddin, Yinyin Wang, Ali Amiryousefi, and Jing Tang. 2020. "Unsupervised Learning and Multipartite Network Models: A Promising Approach for Understanding Traditional Medicine." *Frontiers in Pharmacology* 11: 1319. https://doi.org/10.3389/fphar.2020.01319.

Kononenko, Igor. 2001. "Machine Learning for Medical Diagnosis: History, State of the Art and Perspective." *Artificial Intelligence in Medicine* 23 (1): 89–109. https://doi.org/https://doi.org/10.1016/S0933-3657(01)00077-X.

Kuhn, Max. 2014. "Futility Analysis in the Cross-Validation of Machine Learning Models." https://arxiv.org/abs/1405.6974.

Kuhn, Max, and Hadley Wickham. 2020. *Tidymodels: A Collection of Packages for Modeling and Machine Learning Using Tidyverse Principles.* https://www.tidymodels.org.

Kwan, Janice L, Lisha Lo, Jacob Ferguson, Hanna Goldberg, Juan Pablo Diaz-Martinez, George Tomlinson, Jeremy M Grimshaw, and Kaveh G Shojania. 2020. "Computerised Clinical Decision Support Systems and Absolute Improvements in Care: Meta-Analysis of Controlled Clinical Trials." *Bmj* 370. https://doi.org/10.1136/bmj.m3216.

Lohman, Timothy G., and Alex F. Roche. 1988. *Anthropometric Standardization Reference Manual*. Champaing, IL, USA: Human Kinetics.

Magrabi, Farah, Elske Ammenwerth, Jytte Brender McNair, Nicolet F. De Keizer, Hannele Hyppönen, Pirkko Nykänen, Michael Rigby, et al. 2019. "Artificial Intelligence in Clinical Decision Support: Challenges for Evaluating AI and Practical Implications." *Yearb Med Inform* 28 (01): 128–34.

Martínez-González, Ana AND Toledo, Miguel Angel AND García-Arellano. 2012. "A 14-Item Mediterranean Diet Assessment Tool and Obesity Indexes Among High-Risk Subjects: The PREDIMED Trial." *Plos One* 7 (8): 1–10.

Montani, Stefania, and Manuel Striani. 2019. "Artificial Intelligence in Clinical Decision Support: A Focused Literature Survey." *Yearb Med Inform* 28 (01): 120–27.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Sendak, Mark, Madeleine Clare Elish, Michael Gao, Joseph Futoma, William Ratliff, Marshall Nichols, Armando Bedoya, Suresh Balu, and Cara O'Brien. 2020. "The Human Body Is a Black Box: Supporting Clinical Decision-Making with Deep Learning." In.

Shortliffe, Edward H., and Martin J. Sepúlveda. 2018. "Clinical Decision Support in the Era of Artificial Intelligence." *Jama* 320 (21): 2199–2200. https://doi.org/10.1001/jama.2018.17163.

Siri, WE. 1961. "Techniques for Measuring Body Composition." In, edited by Henzchel A. Brozek J, 224–44. National Academy of Sciences.

Società Italiana di Nutrizione Umana. 2014. *LARN: Livelli Di Assunzione Di Riferimento Di Nutrienti Ed Energia Per La Popolazione Italiana*. Milan, MI, Italy: Società Italiana di Comunicazione Scientifica e Sanitaria.

Spielberger, Charles D. 2010. "State-Trait Anxiety Inventory." *The Corsini Encyclopedia of Psychology*, 1–1.

Stefan, Norbert, Hans-Ulrich Häring, Frank B Hu, and Matthias B Schulze. 2013. "Metabolically Healthy Obesity: Epidemiology, Mechanisms, and Clinical Implications." *The Lancet Diabetes & Endocrinology* 1 (2): 152–62. https://doi.org/https://doi.org/10.1016/S2213-8587(13)70062-7.

V. Weir, J. B. de. 1949. "New Methods for Calculating Metabolic Rate with Special Reference to Protein Metabolism." *The Journal of Physiology* 109 (1-2): 1–9.

Vandenbroeck, Philippe, Jo Goossens, and Marshall Clemens. 2007. "Foresight: Tackling Obesities: Future Choices - Building the Obesity System Map." *PsycEXTRA Dataset*. https://doi.org/10.1037/e602972011-001.

Vidotto, Giulio, Loretta Moroni, Roberto Burro, Luca Filipponi, Gianluigi Balestroni, Ornella Bettinardi, Gisella Bruletti, Ines Giorgi, Marianna Naimo, and Giorgio Bertolotti. 2010. "A Revised Short Version of the Depression Questionnaire." *European Journal of Cardiovascular Prevention and Rehabilitation* 17 (March): 187–97. https://doi.org/10.1097/HJR.0b013e328333edc8.

World Health Organization. 2000. "Obesity: Preventing and Managing the Global Epidemic." World Health Organization.