Deliverable 9.2 (Part C)

Methodological Guidelines for studies to estimate health indicators using linked data and machine learning techniques

October 23, 2020

Submission Date: October 23, 2020

WP9 Lead: Department of Non-Communicable Disease and Injury, Santé Publique France, France
WP9 Co-Lead: Health Information Centre and Institute of Hygiene, Lithuania
Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu|Twitter: @JA_InfAct

This project is co-funded by the Health Programme of the European Union

Table of Contents

Executive summary

Background: The capacity to use data linkage and artificial intelligence to estimate and predict health indicators varies across EU-MSs (European Member States). However, the estimation of health indicators from linked administrative data is challenging due to several reasons such as variability in data sources and data collection methods, availability of a large number of variables, lack of skills and capacity to link and analyze big data. The main objective of this study is to develop the methodological guidelines for studies to guide European countries using linked data and machine learning techniques with new methods/techniques.

Method: We have performed five following steps systematically to develop the methodological guidelines: i. scientific literature review, ii. development of a generic method, iii. identification of inspiring examples from European countries, iv. developing the checklist of guidelines contents and v. the validation of these guidelines by a panel of experts.

Results: We have developed the methodological guidelines, which provide a systematic approach for studies using linked data and machine learning techniques to produce population-based health indicators. These guidelines are validated by a panel of experts including epidemiologists, biostatisticians, data scientists, methodologists, public health professionals and policy experts. These guidelines include a detailed checklist of the following nine items: rationale and objective of the study (i.e., research question), study design, linked data sources, study population/sample size, study outcomes, data extraction and input variables without applying ML-techniques, data preparation to develop and apply an ML-algorithm, data analysis (i.e., statistical techniques, sensitivity analysis and potential issues during data analysis) and common study limitations.

Conclusions: Existing reporting guidelines are not fully designed to capture key methodological aspects applied to studies focused on population health research. This is the first study to develop the methodological guidelines for studies using linked data and machine learning techniques. These guidelines would support researchers to adopt and develop a systematic approach for high-quality research methods. There is a need for high-quality research methods using more linked data and ML-techniques to develop a cross-disciplinary approach for improving population health.

Keywords: Data linkage; Linked data; Machine learning techniques; Artificial intelligence; Guidelines; Methodological guidelines; Statistical techniques; Population health research; and Health indicators

Key points

- Existing reporting guidelines are not fully designed to capture key methodological aspects applied to studies focused on population health research.
-
- There is a need for high-quality research methods using more linked data and ML-techniques to develop a cross-disciplinary approach for improving population health.

I. Introduction

The availability of data generated from different sources is increasing and the possibility to link these data sources with other databases. More efficient ways of data linkage and the use of artificial intelligence (i.e., innovative techniques) are required to generate comparable and timely health information across European countries. Using these innovative techniques have several advantages such as data linkage improves completeness and comprehensiveness of information to guide health policy process[1], and artificial intelligence allows to handle data with a large number of dimensions (features) and units (feature vectors) more efficiently with high precision. Many countries have already invested in the linkage of their traditional health data systems and increased interoperability [2]. The capacity to use data linkage and artificial intelligence (AI) to estimate and predict health indicators varies across EU-MSs (European Member States)[3]. However, the estimation of health indicators from linked administrative data is challenging due to several reasons such as variability in data sources and data collection methods, availability of a large number of variables, lack of skills and capacity to link and analyze big data [4]. Due to varying health information system across MSs, it makes challenging to learn from each other experiences.

To our knowledge, there are no methodological guidelines available, which could systematically guide MSs for using linked data and machine learning techniques (MLTs) to estimate health indicators for population health research. Therefore, the InfAct project has proposed to develop these guidelines, which could guide those MSs who are planning to estimate health indicators using linked data and artificial intelligence (i.e., Machine learning techniques) with new methods/techniques. InfAct (Information for Action)[5] project is a joint action of Member States aiming to develop a more sustainable EU health information system through improving the availability of comparable, robust and policy-relevant health status data and health system performance information. InfAct gathers 40 national health authorities from 28 Member States (MSs).

The main objective of this study is to develop the methodological guidelines for studies to guide European countries to estimate health indicators using linked data and machine learning techniques with new methods/techniques.

II. Methodology

We have performed five following steps systematically to develop the methodological guidelines: i. scientific literature review, ii. development of a generic method, iii. identification of inspiring examples from European countries, iv. developing the checklist of guidelines contents and v. validation of these guidelines by a panel of experts.

i. Literature review

Firstly, we performed a literature search to identify published articles focusing on estimating health indicators using linked data and machine learning techniques on August 1, 2020. We included in our search peer-reviewed methodological articles using both linked data and machine learning techniques in the field of health surveillance and health care performance, related guidelines and systematic reviews that were published in the English language. We excluded those studies published as protocols, scoping reviews or literature reviews, non-methodological studies such as editorials, commentary or perspectives and studies related to life sciences such as RNAi or gene expression. Search strategies are reported in additional file 1. Based on this literature review, we identified various methodological approaches using linked data and machine learning techniques to develop these guidelines.

ii. Generic method

In a previous study[6], we have developed a generic approach using ML (Machine Learning)-algorithm and adopted a supervised machine learning approach. We used this ML-algorithm to predict the incidence of diabetes mellitus using linked data. For this approach, we have defined the following steps to develop an ML-algorithm, which are used to estimate health indicators from linked data: i. selection of final data set, ii. target definition, iii. coding features/variables for a given window of time, iv. split of final data into training and test data sets, v. features/variables selection, vi. training model/algorithm, vii. validation of model/algorithm with test data set and viii. selection of the model/algorithm. This approach can be applied to predict the incidence or prevalence of other health conditions. This generic method provided a systematic approach to develop and apply an ML-algorithm using linked data and to develop these guidelines.

iii. Inspiring examples

In a previous study[7], we have identified 16 studies as inspiring examples from ten European countries that have been performed and some studies are ongoing. We defined inspiring examples as those studies that take into account the use of linked data and/or artificial intelligence (i.e., innovation aspect) to estimate health indicators and implied the related health indicators to target priority public health actions (i.e., surveillance, prevention, promotion, etc.), healthcare strategies or to guide/support public health policies according to their geographical regions. These studies adopted various methodological approaches to estimate health indicators, either by using data linkage (12 studies), or machine learning methods (2 studies) or both data linkage and machine learning approaches (2 studies). These studies were used to develop these guidelines.

iv. Developing the checklist of methodological guidelines contents

Based on the first three steps, we have developed a checklist including the following nine items for guidelines: rationale and objective of the study (i.e., research question), study design, linked data sources, study population/sample size, study outcomes, data extraction and input variables without applying ML-techniques, data preparation to develop and apply an ML-algorithm, data analysis and common study limitations.

v. Validation of methodological guidelines by a panel of experts

The main role of the panel of experts was to review the contents of the guidelines, provide their inputs to improve that and to agree on the final format of the guidelines. The panel experts include epidemiologists, biostatisticians, statisticians, data scientists, methodologists, public health professionals and policy experts. These experts would be working at any of the following departments in the EU Member States: public health institutes, national statistics offices, health information centres, clinical and epidemiological research departments, ministry of health or international health organizations.

Expected outcomes

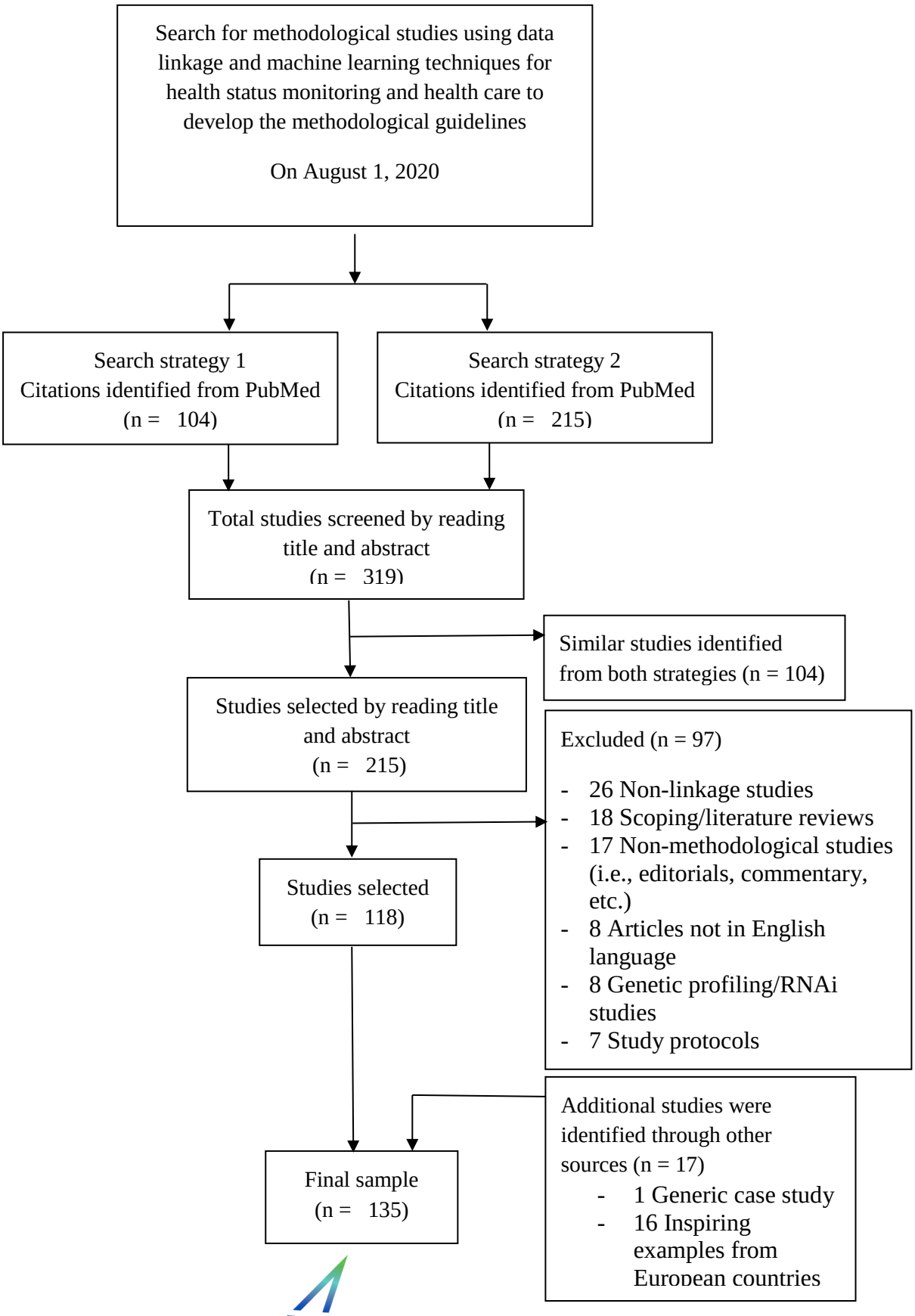
The methodological guidelines are the main outcomes that would guide and support MSs to estimate health indicators using linked data and machine learning techniques.

III. Results

Literature review

We reviewed 215 citations from PubMed and 118 were included in our final sample to develop these methodological guidelines (Fig. 1). 17 additional studies (1 generic method study + 16 inspiring examples) were also included in the final sample. The final sample included 135 studies using either linked data or machine learning techniques to address various research questions either to describing or predicting health indicators in the field of health status monitoring or the evaluation of certain treatments in medical/health care. Among these citations, some guidelines were also identified to adopt the appropriate format of methodological guidelines[8,9].

Fig. 1: Flow diagram of methodological studies using linked data and machine-learning techniques for health status monitoring and health care to develop methodological guidelines



Methodological guidelines for studies using linked data and machine learning techniques

We have developed a checklist including the following nine items for methodological guidelines: rationale and objective of the study (i.e., research question), study design, linked data sources, study population/sample size, study outcomes, data extraction and input variables without applying ML-techniques, data preparation to develop and apply an ML-algorithm, data analysis and common study limitations. Here, we described the rationale of adopting these items with examples of studies under the following domains:

1. Rationale and objective of the study (i.e., research question)

The first step is to define the research question for the proposed study. The PICO criteria (P = Population/patient, I = Intervention/indicator/exposure/risk factor, C = Comparator/alternative intervention [if appropriate], O = Outcome of interest) are used in evidence-based practice to frame and answer clinical and healthcare-related questions [10]. These criteria could be adopted according to population health research questions. The research questions should be simple and smart. The identified research studies focused on common aspects such as estimating the health indicators, associations between health outcomes and exposures, identifying health inequalities, predicting the health indicators/outcomes, classifying population groups to estimating their health outcomes, etc.

2. Study design

The second step is to select the appropriate study design that could best address the proposed research question. In identified research studies, the following were the most commonly used study designs (see additional file 2): cross-sectional studies (for estimating the associations between health outcomes and various exposures); population-based e-cohort (for estimating and predicting health outcomes [e.g. incidence/prevalence] in context of certain risk factors, disease care, classifying population groups to estimating their health outcomes); and a case-control analysis (for comparing health outcomes between cases and controls), etc.

3. Linked data sources

The third step is to select the required linked data sources to answer the proposed research question. The health administrative data sources (i.e., hospital discharge, mortality, primary care/general practitioners, health insurance claims), which are either linked with each other or with other data sources (i.e., disease-specific registries, health surveys, epidemiological cohort studies, vital statistics), are the most commonly used data sources. These data sources are linked using both deterministic and probabilistic data linkage techniques.

4. Study population/sample size

The fourth step is to define the study population according to the proposed research question. Often, the study population is extracted from the national health administrative database linked either with a population-based cohort or disease-specific registry or health survey or with any other administrative database. The linked database allows having a large sample size. The inclusion and exclusion criteria of the study population should be clearly defined according to the research question. The age, sex of the included sample and the period of data collection should be specified.

The sample size is calculated including the standard values of alpha 0.5 and 80% power to detect the potential difference between the two groups.

5. Study outcomes and their estimation at various geographical levels

The fifth step is to define the study outcomes according to the proposed research question. The study outcomes should be clearly defined by taking into account the study population, health condition (to be studied), exposure (intervention or risk factors if relevant) and the defined period of study. The PICO criteria could also be used to define the study outcomes [10].

It is important to estimate the health outcomes at the lowest granularity level (i.e., at the community, metropolitan, departmental or at regional levels) to highlight the variability at the local level and to adopt the health decisions according to the local needs.

6. Data extraction including input variables without applying ML-techniques

This step involves data extraction with required input variables from the linked data sets without applying ML-techniques. The extracted data from linked sources could be exported to a single excel file or a spreadsheet that could be converted to different file formats according to the statistical software to be used for data analysis.

7. Data preparation to develop and apply an ML-algorithm

The seventh step is to prepare the data to develop and apply an ML-algorithm. This step involves the following four sub-steps:

- ***A.*** *Target definition:* First, the targets are identified based on the outcome of interest, and second, these targets are defined either as positive target (cases, for example, pharmacologically treated diabetes patients) or as negative target (controls, for example, non-diabetes patients) for a given time window (e.g., pharmacologically treated diabetes patients in last 6 months as positive targets).
- ***B.*** *Coding and standardization of variables for a given time window:* All the variables, which are common in different linked data sources, are coded

for a given time window (e.g., either 6 or 12 months). After coding, these variables are standardized as mean = 0 and standard deviation = 1.

C. *Split of final data into training and test data sets:* The final data set is split into 80% as a training data set and 20% as a test data set. If there is an imbalance of the number of positive targets (1 group) over the number of negative targets (0 group) in the training dataset, a under-sampling can be performed in the target 0 group to achieve the same number of individuals in both target groups. Later, the selection of features and the models is performed using the training data. The test data is used solely to test the final model performance.

D. *Features/variables selection:* First, all variables with a variance equal to zero are removed, then the ReliefExp score is estimated based on the relevance of each variable to the outcome of interest, to minimize the collinearity effect [11].

8. ***Data analysis:*** The eighth step is the data analysis that includes different statistical techniques, sensitivity/uncertainty analysis and some potential issues that may encounter during the data analysis.

A. ***Statistical techniques used for the estimation of population health indicators:*** Several statistical techniques are applied to linked data either using classical statistical techniques without ML approaches or MLTs. These techniques are used to estimate, classify, predict the population health indicators or to evaluate the health care interventions according to the available linked datasets. A brief description of different techniques is reported in additional file 2.

I. ***Classical statistical techniques without ML approaches:*** Several classical statistical techniques are applied to analyze the linked data set without using ML-techniques. The following are the most commonly used techniques: multilevel linear regression[12], multivariate logistic regression[13], multivariable hierarchical modified Poisson regression[14], Cox regression models[15], LASSO regression[16,17], Generalized Estimating Equation (GEE) models[18], inverse probability weighting methods[19], Blinder-Oaxaca decomposition method[20] and Markov modeling [21].

II. ***ML-techniques:*** Several ML-techniques are applied for studies focused on population health and health care research. Following are the most commonly used supervised ML-techniques: linear and logistic regression, Linear Discriminant Analysis (LDA) model[22,23], partial least square discriminant analysis model [24], decision tree[25], random forest [26] and Gradient Boosting Classifier [GBC] [27,28], k-nearest neighbours/k-means [29],

support vector machine [SVM][30], neural networks[31], convolutional neural networks, hierarchical clustering[32] and XGBoost[33].

To develop and apply ML-techniques, the following are the three main steps to train and select the final model:

a. *Training various models/algorithms:* Most commonly used models are linear discriminant analysis, logistic regression, flexible discriminant analysis and decision trees that are applied to the training data set. The performance of each model is compared in terms of area under the ROC (Receiver Operating Characteristics) curve. ROC curve is an evaluation metric for binary classification problems. It is a probability curve that plots TPR (true positive rates) against FPR (false positive rates) at various threshold values and separates the 'signal' from the 'noise'. The AUC (Area Under the Curve) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve [34]. The higher the AUC, the better the performance of the model at distinguishing between positive and negative classes/targets.
b. *Model validation techniques:* To validate the model, k-fold cross-validation is a commonly used technique. Using this technique, the given data set is split into a *K* number of sections/folds where each fold is used as a testing set at some point. For example, 5-fold cross-validation (K=5) where the data set is split into 5 folds. In the first iteration, the first fold is used to test the model and the rest are used to train the model. In the second iteration, 2nd fold is used as the testing set while the rest serve as the training set. This process is repeated until each fold of the 5 folds has been used as the testing set [35]. This technique allows for estimating the performance or accuracy of the model.
   After the first validation of the models/algorithms using k-fold cross-validation on the training data set, the algorithms' performances are assessed using the testing data set.
c. *Selection of final model/algorithm*: After the model validation, the algorithm selection process is automated by giving the computer-specific metrics including sensitivity, specificity, positive predictive value, negative predictive value, F1-score and kappa. Finally, a single model is retained based on its performance, computational parsimony and its transferability to other databases.

B. ***Sensitivity/uncertainty analysis:*** After the selection of the final model, sensitivity analysis is performed. This analysis refers to identifying the

most influential parameters for a given output of a mathematical computer model (i.e., the sensitivity of output by changing the inputs) or to evaluate the effect of uncertainty in each uncertain computer input variable on a particular model output[36]. It helps to understand the relationship between input and output variables and the robustness of the results of a computing model [37]. The most common methods are variance-based method [38], elementary effects method [39] and regression analysis.

C. ***Potential issues during data analysis:*** During the data analysis, the following are some common issues, which may encounter: missing data, imbalanced datasets and bias-variance tradeoff.

I. ***Missing data:*** In big datasets, missing values are often the main issue that can introduce a substantial amount of bias, make handling and data analysis harder and strongly influence the model performance.

There are three types of missing data [40]: 1. Missing Completely At Random (MCAR): if subjects who have missing data are a random subset of the complete sample of subjects, 2. Missing Not At Random (MNAR): if the probability that an observation is missing depends on information that is not observed, like the value of the observation itself is missing, and 3. Missing At Random (MAR): the probability that an observation is missing commonly depends on information for that subject that is present i.e., the reason for missing data is based on other observed patient characteristics.

Imputations of missing values: Imputation is a process of replacing missing values in a dataset. Following are some common approaches, which could be applied to both types of studies with and without using ML-techniques:

a. ***For analytical methods (non-ML-studies):*** There are three most commonly used techniques i.e., 1. listwise/complete case deletion, 2. single imputation and 3. multiple imputations. Simple/single imputation techniques for handling missing data (such as complete case analysis, overall mean imputation, and the missing-indicator method) produce biased results, whereas multiple imputation techniques yield valid results [40,41].

b. ***For ML-studies:*** There are eight most common ways to replace the missing values in machine learning models: 1. rows/listwise/complete case deletion, 2. replacing with mean/median/mode, 3. assigning a unique category, 4. using most frequent or zero/constant values, 5.

predicting the missing values using linear regression, 6. using algorithms which support missing values, 7. Multivariate Imputation by Chained Equation (MICE) and 8. deep learning (DataWig)[42,43]. Instead of data imputation, a novel method based on additive least square support vector machine (LS-SVM) is potentially a promising technique for tackling missing data in epidemiological studies and community health research [44].

II. ***Imbalanced datasets:*** The second issue is the imbalanced dataset that can skew in class distribution and may bias ML-algorithms. Many ML-techniques, such as neural networks, make more reliable predictions from being trained with balanced data[45]. There are two commonly used approaches to create a balanced data set, first is the under-sampling and the second one is oversampling [45,46].

III. **Bias and variance tradeoff in ML-models:** The third issue is the bias and variance tradeoff. The concept of bias and variance and their relationship with each other is fundamental to the true performance of supervised ML models [47]. Bias refers to the error in the ML-model due to wrong assumptions. A high-bias model will underfit the training data. Variance refers to problems caused due to overfitting. This is a result of the over-sensitivity of the model to small variations in the training data. A model with many degrees of freedom (such as a high-degree polynomial model) is likely to have high variance and thus overfit the training data. Increasing a model's complexity will reduce its bias and increase its variance. This balance is key to finding the most generalizable model [47].

Model tuning/hyperparameter tuning: It is an important step to improve model performance and accuracy. Robust model tuning provides insight on how model structure and hyperparameters influence the model performance[48]. Hyperparameters **a**re adjustable parameters that must be tuned to obtain a model with optimal performance. There are some techniques, which are commonly used to tune the hyperparameters: grid search, random search and Bayesian optimization [49].

9. Common study limitations: Study limitations are important and should be reported to addressing various issues for further research. Different studies using data linkage and ML-techniques reported some common study limitations related

to data sources (linkage, quality, access and privacy), study design and statistical methods. Following are some limitations, which may influence the quality of research studies: ***Data linkage*** (e.g., different data collection methods in different areas make it difficult to link and compare the data, lack of standard methods for data collection); ***Data quality*** (e.g., lacking completeness of information for some routinely collected data sources, unavailability of certain information to improve the results of some analyses, lacking information on secondary cause of death, exclusion of some groups for whom no linkage could be done due to lack of identifier number); ***Access/availability of certain data sources*** (e.g., readily unavailability/inaccessibility of data related to employment, education, occupation and socioeconomic status, lack of data on health inequalities at local levels); ***Data privacy*** (e.g., certain variables cannot be explored due to privacy or confidentiality issues, legal interoperability issues to link various data sources); ***Study design*** (e.g., causality, misclassification of exposure outcome, bias, age of study sample, use of isotropic model of exposure)**;** ***Study methods*** (e.g., overfit or underfit of the model used in ML-studies)

Table 1: Methodological guidelines for studies using linked data and machine learning techniques

Item number	Checklist item	Description
1	**Rationale and objective of the study (i.e., research question)**	Define the rationale and objective of the study by adopting PICO criteria to research studies focused on population health.
2	**Study design**	Select the appropriate study design that could best address the proposed research question.
3	**Linked data sources**	Select the required linked data sources to answer the proposed research question.
4	**Study population**	
4.1		Define the inclusion and exclusion criteria of the study population by taking into account age, sex and period of data collection.
4.2	Sample size calculation	Calculate the sample size including standard values of alpha 0.05 and 80% power to detect the potential difference between the two groups.
5	**Study outcomes**	
5.1		Define the main outcomes by taking into account study population, health condition to be studied, exposure (intervention/risk factors, if relevant) and defined period of study.
5.2	Level of estimation	Estimate health outcomes at the lowest granularity level (i.e., at community, metropolitan, departmental or regional levels).
6	**Data extraction including input**	

	variables without applying ML-techniques	
6.1		Extract data with required input variables from linked data set to a single file or a spreadsheet that could be converted to the required format of the statistical software for data analysis.
6.2	Coding of variables	Code the input variables either as binary or continuous variables for required data analysis.
7	**Data preparation to develop and apply a ML-algorithm**	
7.1		Identify and define the target groups for a given time window based on the outcome of interest.
7.2		Codify all variables for a given time window time and standardize as mean = 0, standard deviation = 1.
7.3		Split of final data set into 80% training and 20% test data set.
7.4		Select features/variables after the removal of all variables with a variance equal to zero.
7.5		Estimate the RelifExp score based on the relevance of each variable to differential between +ve and –ve targets.
8	**Data analysis**	
	A. Statistical techniques	
8.1	I. Classical statistical techniques	Select an appropriate statistical technique to address the proposed research question according to the study objectives and the available data.
	II. ML-techniques	
8.2		Train various models and compare the performances of each model in terms of area under the ROC curve.
8.3		Validate the model performance using k-fold cross-validation first on training data set, and then assess the model performance on test data set.
8.4		Select the final model based on specific performance metrics including sensitivity, specificity, PPV*, NPV*, F1-score and kappa.
	B. Sensitivity/uncertainty analysis	
8.5		Perform a sensitivity analysis to identify the most influential parameters for a given output of a model.
8.6		Select an appropriate method to perform the sensitivity analysis.
8.7		Calculate the uncertainty in estimates using 95% CI* and describe the source of uncertainty (if relevant).
	C. Potential issues during data analysis	
	I. Missing data	
8.8		Identify the missing data in the given dataset.
8.9		Apply an appropriate technique for the imputation of missing values in the given data set.
8.10	II. Imbalanced target group in a given dataset	Apply an appropriate technique to create a balanced data set either using under-sampling or oversampling approach.
8.11	III. Bias and variance tradeoff	Find the most generalizable model to keep the balance between bias and variance.

9	**Study limitations**	Describe the study limitations related to data sources (i.e., linkage, quality, access and privacy), study design, study population and statistical method used (if relevant).

**PPV: Positive Predictive Value, NPV: Negative Predictive Value, CI: Confidence interval*

Validation of guidelines by the panel of experts

The first draft of these guidelines was shared with the panel of experts to review the contents and the format. The guidelines were updated according to their inputs until a consensus was reached for the final version.

IV. Discussion

Main results: We have developed the methodological guidelines, which provide a systematic approach for studies using linked data and machine learning techniques to produce population-based health indicators. These guidelines include a checklist of the following nine items: rationale and objective of the study (i.e., research question), study design, linked data sources, study population/sample size, study outcomes, data extraction including input variables without applying ML-techniques, data preparation to develop and apply an ML-algorithm, data analysis (i.e., statistical techniques, sensitivity analysis and potential issues during data analysis) and common study limitations.

There are few studies available, which describe the reporting guidelines for linked data focused on population health research. The first study that illustrates the guidelines to evaluate the methodological quality of studies using linked data and to report their results in a consistent, high-quality manner [9]. The second study defines the best reporting practices as guidelines for accurate and transparent reporting of health estimates for studies that calculate health estimates for multiple populations (in time or space) using multiple information sources[8]. Another study developed TIDieR-PHP (Template for Intervention Description and Replication-Population Health and Policy) checklist to improve the reporting of PHP interventions[50]. These guidelines are important for reporting of key characteristics of studies in general. Nevertheless, the existing reporting guidelines are not fully designed to capture key methodological aspects applied to studies focused on population health research.

Scope: These guidelines define a systematic approach for studies using linked data and ML-techniques for population health research. We used peer-reviewed published methodological studies, which applied data linkage and ML-approaches in the field of health status monitoring and medical/health care for the estimation and prediction of health indicators. These guidelines offer a general framework for using linked data and ML-techniques and are flexible enough to integrate new methods used for population health research.

Implications: These guidelines would assist public health researchers and epidemiologists to develop and adopt new methods/techniques using linked data and machine learning approaches for their studies. Moreover, these would add to high-quality evidence-based research to guide health policy decisions.

Strengths and limitations: This is the first study to develop methodological guidelines for studies using data linkage and machine learning techniques for improving the quality of research methods for population health research. A large group of the panel of experts from various disciplines provided their inputs, which improved the quality of these guidelines from different perspectives.

There are very few limitations: first, we provided a systematic approach with some general and basic techniques that are most commonly applied for studies using data linkage and ML-techniques. More techniques are possible, which are not reported here and could be applied to answer various research questions to improve the population health research. Second, there are more studies possible, which have applied these techniques and are not reported in this study. We provided at least one example of each statistical technique to better understand the method.

Recommendations: We proposed the following recommendations that could not only address some of the study limitations identified but also promote the research studies using linked data and ML-techniques:

Data sources: data related to employment, education, occupation and socioeconomic status should be readily available/accessible for analysis related to the health status, standard methods for data collection should be implemented in a health information system and routinely data collected from various administrative sources should improve their quality concerning to the completeness of the information. *Data regulations:* specific mandates to ensure data availability/access/capture and safe storage should be an integral part of a national/regional health information system, differences in the implementation and interpretation of the EU-GDPR (General Data Protection Regulations) and additional national regulations should be mapped and if possible harmonize the implementation of GDPR across EU-MSs [51]. *Study design:* the rational selection of the study design using linked data is important to avoid certain methodological limitations. *Statistical methods:* the use of an appropriate statistical model is important to have more reliable results. *Knowledge translation*: better approaches to translate estimated health indicators into health policy are required. *Collaborations:* more collaborations among the Member States for an exchange of inspiring examples/best practices in using linked data and machine-learning approaches are needed in the future among European countries and joint country studies on using machine-learning techniques for public health research are needed.

V. Conclusions

Existing reporting guidelines are not fully designed to capture key methodological aspects applied to studies focused on population health research. This is the first study to develop the methodological guidelines for studies using linked data and machine learning techniques. These guidelines would support researchers to adopt a systematic approach with high-quality research methods. Using linked data and ML-techniques have the potential to add value in research focused on population health. However, the overall generalizability of ML-models in real-world data is critical and the researchers should aware of their data limitations. There is a need for high-quality research methods using more linked data and ML-techniques to develop a cross-disciplinary approach for improving population health.

VI. List of abbreviations

EU: European Union

MSs: Member States

AI: Artificial Intelligence

MLTs: Machine Learning Techniques

InfAct: Information for Action i.e., a joint action of Member States to establish a sustainable European health information system

PICO Criteria: Population-Intervention-Comparator-Outcome Criteria

LASSO: Least Absolute Shrinkage and Selection Operator

GEE: Generalized Estimating Equation

GBC: Gradient Boosting Classifier

SVM: Support Vector Machine

LDA: Linear Discriminant Analysis model

FDA: Flexible Discriminant Analysis model

XGBoost: Extreme Gradient Boosting

HWNNs: Hybrid Wavelet Neural Networks

SOM: Self-Organizing Maps

ROC: Receiver Operating Characteristics

MCAR: Missing Completely At Random

MNAR: Missing Not At Random

MAR: Missing At Random

MICE: Multivariate Imputation by Chain Equation

LS-SVM: Lease Square-Support Vector Machine

PPV: Positive Predictive Value

NPV: Negative Predictive Value

GDPR: General Data Protection Regulations

Declarations

Ethics approval and consent to participate
Not applicable

Consent for publication
All authors gave the consent for publication.

Availability of data and materials
Not applicable

Competing interests
All other authors have declared that they have no competing interests related to the work.

Funding
This research has been carried out in the context of the project '801553 / InfAct', which has received funding from the European Union's Health Programme (2014-2020).

Acknowledgments
We acknowledge the great support from InfAct partners who provided their inputs to improve the quality of this document:
Stefan Mathis-Edenhofer, Claudia Hable (Austrian National Public Health Institute GÖG); Herman Van Oyen (Sciensano, Belgium); Jelena Dimnjakovic, Jakov Vukovic, Ivan Pristas (National Institute of public health, division of health informatics and biostatistics, Croatia); Májek Ondřej (National Institute of Health Information and Statistics); Hanna Tolonen (THL, Department of Public Health Solutions, Finland); Romana Haneef, Anne Gallay, (Department of Non-Communicable Diseases and Injuries, Santé Publique France, Saint-Maurice); Rodolphe Thiebaut (University of Bordeaux, France); Mariken J. Tijhuis (National Institute for Public Health and the Environment [RIVM], The Netherlands), Rok Hrzic (Department of International Health, Care and Public Health Research Institute - CAPHRI, University[2] of Maastricht University, Maastricht, The Netherlands); Unim Brigid, Luigi Palmieri (Department of cardiovascular, Endocrine-metabolic Diseases and Aging,Italy); Ausra Zelviene, Rita Gaidelyte (Institute of Hygiene, Department of Health information Centre and

Health Statistics, Lithuania); Tina Lesnik, Metka Zaletel, Tatjana Kofol Bric (NIJZ, Slovenia); Isabel Noguer, Rodrigo Sarmiento, Alicia Padron Monedero (ISCIII, Spain).

VII. References

1. Lloyd K, McGregor J, John A, et al. A national population-based e-cohort of people with psychosis (PsyCymru) linking prospectively ascertained phenotypically rich and genetic data to routinely collected records: Overview, recruitment and linkage. *Schizophrenia Research*. 2015;166(1):131-136.
2. Delnord M, Szamotulska K, Hindori-Mohangoo AD, et al. Linking databases on perinatal health: a review of the literature and current practices in Europe. *Eur J Public Health*. 2016;26(3):422-430.
3. Haneef R, Delnord M, Vernay M, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. *Archives of Public Health*. 2020;78(1):55.
4. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. *Health Services Research*. 2010;45(5p2):1468-1488.
5. Joint Action on Health Information: https://www.inf-act.eu/. 2018.
6. Haneef R, Fuentes S, Fosse-Edorh S, et al. Use of Artificial Intelligence for Public Health Surveillance: A case study to develop a Machine Learning-algorithm to predict the incidence of Diabetes Mellitus (manuscript in preparation)2020.
7. InfAct R. *Inspiring Examples from European Countries*. 2020.
8. Stevens G, Alkema L, Black R, et al. Guidelines for Accurate and Transparent Health Estimates Reporting: the GATHER statement. *Lancet*. 2016;388(10062):e19-e23.
9. Bohensky M, Jolley D, Sundararajan V, Evans S, Ibrahim J, Brand C. Development and validation of reporting guidelines for studies involving data linkage. *Aust N Z J Public Health*. 2011;35(5):486-489.
10. Illinois Uo. What is a PICO model?:https://researchguides.uic.edu/c.php?g=252338&p=3954402. 2020.
11. Robnik-Sikonja M, Kononenko I. An adaptation of Relief for attribute estimation in regression: http://www.clopinet.com/isabelle/Projects/reading/robnik97-icml.pdf. 1997.
12. Mason KE, Pearce N, Cummins S. Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank. *Lancet Public Health*. 2018;3(1):e24-e33.
13. Sultan A, West J, Grainge M, et al. Development and validation of risk prediction model for venous thromboembolism in postpartum women: multinational cohort study. *Bmj*. 2016;5(355).
14. Patel K, Spertus J, Khariton Y, Tang Y, Curtis L, Chan P. Association Between Prompt Defibrillation and Epinephrine Treatment With Long-Term Survival After In-Hospital Cardiac Arrest. *Circulation*. 2018;137(19):2041-2051.
15. Fogg AJ, Welsh J, Banks E, Abhayaratna W, Korda RJ. Variation in cardiovascular disease care: an Australian cohort study on sex differences in receipt of coronary procedures. *BMJ Open*. 2019;9(7):e026507.
16. Odgers D, Tellis N, Hall H, Dumontier M. Using LASSO Regression to Predict Rheumatoid Arthritis Treatment Efficacy. *AMIA Jt Summits Transl Sci Proc*. 2016;20:176-183.
17. Orriols L, Avalos-Fernandez M, Moore N, et al. Long-term chronic diseases and crash responsibility: a record linkage study. *Accid Anal Prev*. 2014;71:137-143.
18. Patte K, Laxer R, Qian W, Leatherdale S. An analysis of weight perception and physical activity and dietary behaviours among youth in the COMPASS study. *SSM Popul Health*. 2016;2:841-849.

19. Astley CM, Chew DP, Keech W, et al. The Impact of Cardiac Rehabilitation and Secondary Prevention Programs on 12-Month Clinical Outcomes: A Linked Data Analysis. *Heart Lung Circ*. 2020;29(3):475-482.
20. Van der Heyden J, Van Oyen H, Berger N, De Bacquer D, Van Herck K. Activity limitations predict health care expenditures in the general population in Belgium. *BMC Public Health*. 2015;15:267.
21. Asaria M, Walker S, Palmer S, et al. Using electronic health records to predict costs and outcomes in stable coronary artery disease. *Heart*. 2016;102(10):755-762.
22. Ezzati A, Zammit AR, Harvey DJ, et al. Optimizing Machine Learning Methods to Improve Predictive Models of Alzheimer's Disease. *Journal of Alzheimer's Disease*. 2019;71:1027-1036.
23. Yang T, Zhang L, Yi L, et al. Ensemble Learning Models Based on Noninvasive Features for Type 2 Diabetes Screening: Model Development and Validation. *JMIR Med Inform*. 2020;8(6):e15431-e15431.
24. Tuti T, Agweyu A, Mwaniki P, Peek N, English M. An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from Kenya. *BMC Med*. 2017;15(1):201.
25. Goldstein S, Zhang F, Thomas J, Butryn M, Herbert J, Forman E. Application of Machine Learning to Predict Dietary Lapses During Weight Loss. *J Diabetes Sci Technol*. 2018;12(5):1045-1052.
26. Montazeri M, Montazeri M, Montazeri M, Beigzadeh A. Machine learning models in breast cancer survival prediction. *Technol Health Care*. 2016;24(1):31-42.
27. Rahimian F, Salimi-Khorshidi G, Payberah AH, et al. Predicting the risk of emergency admission with machine learning: Development and validation using linked electronic health records. *PLoS Med*. 2018;15(11):e1002695.
28. Zhang L, Wang Y, Niu M, Wang C, Wang Z. Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study. *Scientific Reports*. 2020;10(1):4406.
29. Zhao M, Tang Y, Kim H, Hasegawa K. Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer. *Cancer Inform*. 2018;17:1176935118810215-1176935118810215.
30. Zheng T, Xie W, Xu L, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics*. 2017;97:120-127.
31. Seligman B, Tuljapurkar S, Rehkopf D. Machine learning approaches to the social determinants of health in the health and retirement study. *SSM - Population Health*. 2018;4:95-99.
32. Ahlqvist E, Storm P, Käräjämäki A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. *The Lancet Diabetes & Endocrinology*. 2018;6(5):361-369.
33. Maeta K, Nishiyama Y, Fujibayashi K, et al. Prediction of Glucose Metabolism Disorder Risk Using a Machine Learning Algorithm: Pilot Study. *JMIR Diabetes*. 2018;3(4):10212.
34. Aniruddha BHANDARI. AUC-ROC Curve in Machine Learning: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/. 2020.
35. MUJTABA H. What is Cross Validation in Machine Learning?: https://www.mygreatlearning.com/blog/cross-validation/. 2020.
36. Introduction to Sensitivity Analysis. In: *Global Sensitivity Analysis. The Primer*.1-51.
37. Sensitivity Analysis: From Theory to Practice. In: *Global Sensitivity Analysis. The Primer*.237-275.
38. Variance-Based Methods. In: *Global Sensitivity Analysis. The Primer*.155-182.
39. Elementary Effects Method. In: *Global Sensitivity Analysis. The Primer*.109-154.
40. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: A gentle introduction to imputation of missing values. *Journal of Clinical Epidemiology*. 2006;59(10):1087-1091.
41. Chinomona A, Mwambi H. Multiple imputation for non-response when estimating HIV prevalence using survey data. *BMC Public Health*. 2015;15(1):1059.
42. Maladkar K. 5 Ways To Handle Missing Values In Machine Learning Datasets: https://analyticsindiamag.com/5-ways-handle-missing-values-machine-learning-datasets/. 2018.

43. Badr W. 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples): https://towardsdatascience.com/6-different-ways-to-compensate-for-missing-values-data-imputation-with-examples-6022d9ca0779. 2019.
44. Wang G, Deng Z, Choi KS. Tackling Missing Data in Community Health Studies Using Additive LS-SVM Classifier. *IEEE J Biomed Health Inform*. 2018;22(2):579-587.
45. Shelke MS, Deshmukh PR, Shandilya VK. A Review on Imbalanced Data Handling Using Undersampling and Oversampling Technique: https://www.ijrter.com/papers/volume-3/issue-4/a-review-on-imbalanced-data-handling-using-undersampling-and-oversampling-technique.pdf. 2017.
46. Brownlee J. Random Oversampling and Undersampling for Imbalanced Classification: https://machinelearningmastery.com/random-oversampling-and-undersampling-for-imbalanced-classification/. 2020.
47. Rashidi HH, Tran NK, Betts EV, Howell LP, Green R. Artificial Intelligence and Machine Learning in Pathology: The Present Landscape of Supervised Methods. *Acad Pathol*. 2019;6:2374289519873088.
48. Glushkovsky A. Robust Tuning for Machine Learning: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/1868-2018.pdf 2018.
49. Jordan J. Hyperparameters tuning: https://www.jeremyjordan.me/hyperparameter-tuning/. 2017.
50. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. *BMJ*. 2018;361:k1079.
51. EPRS. How the General Data Protection Regulation changes the rules for scientific research: https://www.europarl.europa.eu/RegData/etudes/STUD/2019/634447/EPRS_STU(2019)634447_EN.pdf. 2019.

VIII. Appendices

Additional file 1: Search strategies

Search strategy 1: ((Linked data [Title/Abstract] OR Machine learning techniques [Title/Abstract]) AND Guidelines [Title/Abstract]))

Search strategy 2: ((Health indicators [Title/Abstract] OR Linked data [Title/Abstract]) OR Machine learning techniques [Title/Abstract]) AND Guidelines [Title/Abstract]))

Additional file 2: Brief description of statistical techniques used in various studies

We have identified 19 different statistical techniques used in various studies either for health monitoring or to improve medical or health care. More techniques are possible to apply. However, here we describe the brief description of some models used in various studies under two categories: 1. Classical statistical techniques and 2. Machine learning techniques.

1. Classical statistical techniques (without applying machine-learning techniques) (N = 9)

i. Multilevel, multiple linear regression models

Model description: Multiple linear regression (MLR), also known simply as multiple regression, is a statistical technique that uses several explanatory variables to predict the outcome of a response variable. The goal of multiple linear regression (MLR) is to model the linear relationship between the explanatory (independent) variables and response (dependent) variable.
Title of the study: Associations between fast food and physical activity environments and adiposity in mid-life: cross-sectional, observational evidence from UK Biobank
Link to the study: https://doi.org/10.1016/S2468-2667(17)30212-8
Study design: Cross-sectional study
Domain: Health status monitoring
Data sources used: Population based cohort linked with spatial datasets including information on physical environment.
Use of model to determine: To examined whether neighbourhood exposure to fast-food outlets and physical activity facilities were associated with adiposity in UK adults
Type of model: Regression model
Models and parameters used in the study: Multilevel, multiple linear regression models with random intercepts and random coefficients were used for the main exposure to estimate independent associations between each environmental exposure and each adiposity outcome, accounting for the nesting of individuals within assessment centres. Initially the model was adjusted only for age and sex (model 0), then for likely demographic confounders (age, sex, ethnicity, area deprivation, and urbanicity; model 1), then further adjusted for individual level socioeconomic characteristics (income, education, and employment status; model 2) and, finally, for the non-exposure environmental feature (proximity to fast food or density of physical activity facilities) and neighbourhood residential density (model 3). As well as adjusting for potential confounding by sex and income, the models were also tested fully adjusted for effect modification by these variables. The results were reported in stratified form where models with interaction terms for sex or income were statistically different from those without (likelihood ratio test $p<0.05$). This study also estimated the same models using height as a negative control outcome.

ii. Multivariate logistic regression

Model description: Multiple logistic regression is distinguished from multiple linear regression in that the outcome variable (dependent variables) is dichotomous (e.g., diseased or not diseased). Its aim is the same as that of all model-building techniques: to derive the best-fitting, most parsimonious (smallest or most efficient), and biologically reasonable model to describe the relationship between an outcome and a set of predictors. Here, the independent variables are called *covariates*. Importantly, in multiple logistic regression, the predictor variables may be of any data level (categorical, ordinal, or continuous). A major use of this technique is to examine a series of predictor variables to determine those that best predict a certain outcome.

Title of the study: Development and validation of risk prediction model for venous thromboembolism in postpartum women: multinational cohort study
Link to the study: https://doi.org/10.1136/bmj.i6253
Study design: Cohort study
Domain: Medical care
Data sources used: The Clinical Practice Research Datalink (CPRD) is a large, longitudinal UK primary care linked database that covers 6% of the population was used as a derivation cohort and Swedish birth registry as a validation cohort.
Use of model to determine: To develop and validate a risk prediction model for venous thromboembolism in the first six weeks after delivery (early postpartum).
Type of model: Classification model
Models and parameters used in the study: The occurrence of venous thromboembolism during the first six weeks postpartum was treated as a binary outcome measure. For each of the 22 candidate predictors, we used a univariable logistic regression model to calculate the unadjusted odds ratio. For derivation of the risk prediction model, initially all candidate predictors in a multivariable logistic regression model were included. A clustering term was fitted to take account of consecutive pregnancies within women during the study period and used fractional polynomials to model potential non-linear relations between outcome and continuous predictors. All variables are coded as binary (0 or 1 for absence or presence of a risk factor), except for age, body mass index (BMI), and birth weight. These three variables were transformed on the basis of fractional polynomial regression (first degree) analysis. The value −9.103 is the intercept, and other numbers are the estimated regression coefficients for the predictors, which indicate their mutually adjusted relative contribution to the outcome risk. The regression coefficients represent the log odds ratio for a change of 1 unit in the corresponding predictor. The predicted risk of VTE=1/1+e−riskscore.

iii. Multivariable hierarchical modified Poisson regression

Model description: Modified Poisson regression, which combines a log Poisson regression model with robust variance estimation, is a useful alternative to log binomial regression for estimating relative risks.
Title of the study: Association Between Prompt Defibrillation and Epinephrine Treatment With Long-Term Survival After In-Hospital Cardiac Arrest
Link to the study: https://doi.org/10.1161/CIRCULATIONAHA.117.030488
Study design: Cohort study
Domain: Medical care
Data sources used: Data from American Heart Association's GWTG (Get With The Guidelines)-Resuscitation registry, which is a large, prospective, quality-improvement registry of IHCAs (In-Hospital Cardiac Arrests) linked with inpatient files of Medicare.

Use of model to determine: To examine long-term survival according to the promptness of defibrillation and epinephrine administration in patients with an IHCA resulting from shockable and nonshockable rhythms, respectively.
Type of model: Regression model
Models and parameters used in the study: To assess the associations between prompt treatment and long-term survival for each rhythm type, hierarchical multivariable modified Poisson regression models were constructed. Modified Poisson regression was used to correct for overestimation of estimates of effect observed with odds ratios when the outcome rate exceeds 10%.Instead, Poisson models yield relative risk estimates obtained from a Poisson distribution. Moreover, these models were hierarchical models, with site as a random effect and patient-level factors as fixed effects. Specifically, they modeled as fixed effects age, sex, race, time to start of cardiopulmonary resuscitation, location of cardiac arrest, and different coexisting conditions and events present within 24 hours before the cardiac arrest. In addition, the models were adjusted for interventions in place at the time of cardiac arrest, day of the week, and calendar of year admission of cardiac arrest.

iv. Cox regression model

Model description: The Cox proportional-hazards model is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables.
Title of the study: Variation in cardiovascular disease care: an Australian cohort study on sex differences in receipt of coronary procedures
Link to the study: http://dx.doi.org/10.1136/bmjopen-2018-026507
Domain: Medical care
Study design: Prospective cohort study
Data sources used: A population cohort data linked with hospital data and a death registry.
Use of model to determine: Time dependent variation
Type of model: Regression model
Models and parameters used in the study: Cox proportional hazard regression was used to model the association between sex and receipt of coronary procedures. For each analysis, participants contributed person-years from the date of index admission for AMI or angina until either the specified outcome of interest, death from any cause or end of follow-up (30 June 2016), whichever was the earliest, to a maximum of one calendar year. Data from patients in the angina sample were also censored if they were subsequently admitted with AMI. Proportional hazards assumption was tested, with the p-value set a priori to $p<0.01$. All analyses were conducted separately for patients whose index admission was for AMI, and for those whose index admission was for angina. Patients presenting concurrently with AMI and angina were included in the AMI sample.
For each outcome, we calculated crude incidence rates separately for men and women, then ran a series of Cox regression models to estimate HRs in relation to sex. Model 1 was adjusted for age (5-year age categories from 45 to 54 years through to ≥80 years). Model 2 was adjusted for age and sociodemographic variables (country of birth, region of residence, highest qualification, private health insurance and marital status). Model 3 was further adjusted for additional baseline health characteristics (obesity, physical functioning and psychological distress). Participants with missing values for covariates were included in the models, with missing coded as a separate category.

v. LASSO (Least Absolute Shrinkage and Selection Operator) model

Model description: LASSO model is a regression analysis method that performs both variable selection and regularization in order to enhance the prediction accuracy and interpretability of the statistical model it produces.

Title of the study: Long-term chronic diseases and crash responsibility: A record linkage study
Link to the study: https://doi.org/10.1016/j.aap.2014.05.001
Domain: Health status monitoring
Study design: A case-control study
Data sources used: Data from three French national databases were extracted and matched: the national healthcare insurance database, police reports and the national police database of injurious crashes.
Use of model to determine: To assess the population impact of chronic conditions on the risk of road traffic crashes
Type of model: Regression model
Models and parameters used in the study: A single model adjusted for crash-related and socio-demographic factors, including all the 299 long-term diseases as covariates, using the Lasso (least absolute shrinkage and selection operator) method was fitted (Avalos et al., 2012; Tibshirani, 1996). Adjustment variables (age, gender, socioeconomic category, year, season, day, time and location of crash, vehicle type, injury severity, blood alcohol concentration and exposure to level 2 and 3 medicines)were forced into the model; the proper amount of shrinkage of the long-term disease covariates was estimated using the Akaike information criterion (AIC) and was corrected for bias. One limitation of the Lasso method is that with a proper amount of shrinkage relevant covariates are retained, but so too are a few additional irrelevant ones (though, typically, their estimates are small). Different procedures have been proposed in the literature to address this particular problem, such as those based on bootstrap-enhanced Lasso (Avaloset al., 2012; Bach, 2008; Bunea et al., 2011). Thus, to reduce the false discovery rate, only chronic condition covariates chosen more frequently by the Lasso over the 5000 bootstrap samples were selected and investigated further. The threshold frequency (75% of the bootstrapped models) was also chosen by AIC. In order to control for multiple medical conditions, a variable was introduced in the multivariable analyses, representing all other chronic diseases than the ones specifically identified. We used the R package glmnet (R Development Core Team, 2011). We also fitted 299 separate logistic regression models, disease by disease, using conventional maximum likelihood adjusting for crash-related and socio-demographic factors. Analyses were performed with and without Bonferroni correction for multiple testing, in order to compare results from the Lasso method with a conventional modeling strategy.

vi. Generalized Estimating Equation (GEE) models

Model description: GEE is used to estimate the parameters of a generalized linear model with a possible unknown correlation between outcomes. Parameter estimates from the GEE are consistent even when the covariance structure is misspecified, under mild regularity conditions. The focus of the GEE is on estimating the average response over the population ("population-averaged" effects) rather than the regression parameters that would enable prediction of the effect of changing one or more covariates on a given individual.
Title of the study: An analysis of weight perception and physical activity and dietary behaviours among youth in the COMPASS study
Link to the study: https://doi.org/10.1016/j.ssmph.2016.10.016
Domain: Health status monitoring
Study design: Cohort study
Data sources used: This study used 2-year linked data of 19,322 grade 9-12 students from Year 2 (Y_2:2013-2014) and 3 (Y_3:2014-2015) of the COMPASS study.
Use of model to determine: to examine how weight perception influences physical activity (PA) and diet among youth.
Models and parameters used in the study: Generalized Estimating Equations (GEE) models were used to test the effect of Y3 weight perception (underweight, overweight, “about right”) on the various

Y3 outcome measures of PA and dietary behaviours, adjusting for Y3 covariates (grade, race/ethnicity, weekly spending money, school area median household income) and the Y2 outcome health behaviour. Models were stratified by gender and BMI status. In other words, the models included Y3 data for the predictor, covariate, and outcome measures, and adjusted for the outcome measure from Y2 data, in order to strengthen inferences. The GEE model is an extension of generalized linear models to correlated data, simply modelling the mean response and treating covariance as nuisance. It produces consistent estimates for regression parameters and can be used for continuous, categorical (including binary), and ordinal measurements. In our analyses, we specified identity link function for continuous outcomes, logit for binary outcomes and cumulative logit for ordinal outcomes. Schools were included in the models as clusters to take account of within-school correlation. Squared root transformation was used for continuous outcome variables to meet model assumptions.

vii. Inverse probability weighting (IPW) methods

Model description: Inverse probability weighting is a statistical technique for calculating statistics standardized to a pseudo-population different from that in which the data was collected. Study designs with a disparate sampling population and population of target inference (target population) are common in application.
Title of the study: The Impact of Cardiac Rehabilitation (CR) and Secondary Prevention Programs on 12-Month Clinical Outcomes: A Linked Data Analysis
Link to the study: https://doi.org/10.1016/j.hlc.2019.03.015
Domain: Medical care
Study design: Retrospective cohort design
Data sources used: Cardiac Rehabilitation databases were linked to hospital administrative datasets
Use of model to determine: to determine if CR attendance impacts on cardiovascular readmission, morbidity and mortality.
Models and parameters used in the study: An inverse probability weighting (IPW) model was used to account for selection bias and unequal probabilities of those patients attending or not attending CR. Factors in the IPW model included age, gender, primary diagnosis, Charlson Index, prior HF, coronary disease, AF, revascularization, malignancy and social factors measured by the IRSAD. The IRSAD measures high and low income, degree of house mortgage, size of home, educational level, qualifications or none, employed as a professional, a manager, low skilled worker, machinery operator or labourer, high rent, number of cars or none, overcrowding, divorced, low rent, disability, unemployed single parent family, no internet access and jobless parents from the Australian Census. Each ED presentation resulting in a separation was considered a single hospitalization and each cardiac admission was counted as a single separation, with admissions involving transfer(s) merged as one. Readmission within 24 hours was not counted as a new event. For assessment of associations between CR attendance and cardiovascular events (cardiovascular readmission, death, new/re-MI, HF, AF or stroke), an analysis confined to those patients referred to CR was undertaken. The balance of the IPW weighted population is presented in Table 1. Further, cardiovascular events occurring prior to 70 days from discharge were removed from the analysis and outcomes were measured post CR program within 12 months. Associations with CR attendance and outcomes were measured by Cox proportional hazard models applied to the IPW population and stratified by primary cardiac diagnosis, referring hospital, Charlson Index and adjusted for age, gender and socioeconomic status. The proportional hazards assumption was assessed and found to be valid.

viii. Blinder-Oaxaca decomposition method

Model description: The Blinder–Oaxaca decomposition is a statistical method that explains the difference in the means of a dependent variable between two groups by decomposing the gap into that part that is due to differences in the mean values of the independent variable within the groups, on the one hand, and group differences in the effects of the independent variable, on the other hand.
Title of the study: Activity limitations predict health care expenditures in the general population in Belgium
Link to the study: https://doi.org/10.1186/s12889-015-1607-7
Study design: Retrospective cohort design
Domain: Health status monitoring
Data sources used: Data from the Belgian Health Interview Survey 2008 were linked with data from the compulsory national health insurance (n = 7,286).
Use of model to determine: The predictive value of the GALI (Global Activity Limitation Indicator) on health care expenditures in relation to the presence of chronic conditions.
Models and parameters used in the study: To study the factors contributing to the difference in health expenditure between persons with and without activity limitations, the Blinder-Oaxaca decomposition method was used. Although multivariate regression models are suitable to address differences in the importance of individual factors, the Blinder-Oaxaca technique demonstrates the relative importance of each predictor. The decomposition illustrates the fraction of the gap in health care expenditures that is attributable to group differences in the magnitude of the determinants (the explained or prevalence component) and to group differences in the effects of these determinants (the unexplained or impact component). The Blinder-Oaxaca decomposition method is particularly useful to study differences in health care expenditures between two groups, but it has also been used in studies in which the contribution of both the prevalence and the impact of determinants to explain differences between groups was investigated for other health outcomes.

ix. Markov modelling

Model description: In probability theory, a Markov model is a stochastic model used to model randomly changing systems. It is assumed that future states depend only on the current state, not on the events that occurred before it (that is, it assumes the Markov property). Generally, this assumption enables reasoning and computation with the model that would otherwise be intractable.
Title of the study: Electronic health records
Link to the study: http://dx.doi.org/10.1136/heartjnl-2015-308850
Domain: Medical care
Study design: Retrospective cohort design
Data sources used: The analysis was based on 94 966 patients with stable-CAD (Coronary Artery Disease) in England between 2001 and 2010, identified in four prospectively collected, linked EHR sources.
Use of model to determine: To predict lifetime costs and health outcomes of patients with stable coronary artery disease (stable-CAD) stratified by their risk of future cardiovascular events, and to evaluate the cost-effectiveness of treatments targeted at these populations.
Type of model: Predictive model
Models and parameters used in the study: A state transition model (shown in figure 1) was developed to capture the natural history of patients with stable-CAD. The structure of the model was determined with reference to both previous models in CVD13 and expert clinical advice. All patients entered the model in the stable-CAD state and progressed through the model until they experienced either CVD or non-CVD mortality. The time horizon of the model was, therefore, the patient's remaining lifetime. The model captured time varying and age-dependent risks, costs and health-related quality of life (HRQoL) in 90-day segments. Costs and HRQoL were attached to model states

and, in order to stratify by patients' baseline risk, adjusted for patient covariates at baseline as well as for age and for time elapsed following non-fatal events. Model predicted costs, life years and QALYs were discounted at 3.5% per annum in keeping with the guidelines in England.18 While only first occurrences of non-fatal CVD events were explicitly modelled, further non-fatal events were implicitly captured in the time varying risk, cost and HRQoL estimates.

2. Supervised Machine Learning Techniques (N = 10)

i. Linear Discriminant Analysis (LDA)

Model description: Linear discriminant analysis (LDA), normal discriminant analysis (NDA), or discriminant function analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition, and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events. The resulting combination may be used as a linear classifier, or, more commonly, for dimensionality reduction before later classification.
Title of the study: Optimizing Machine Learning Methods to Improve Predictive Models of Alzheimer's Disease
Link to the study: https://dx.doi.org/10.3233/JAD-190262
Study design: cohort study
Domain: Clinical care
Data sources used: Data from an ongoing cohort study
Use of model to determine: To classify cognitively normal (CN) individuals from Alzheimer's disease (AD) and to predict longitudinal outcome in participants with mild cognitive impairment (MCI
Type of model: Predictive model
Models and parameters used in the study: In this study, four features set and six machine-learning methods (decision trees, support vector machines, K-nearest neighbor, ensemble linear discriminant, boosted trees, and random forests) were used to classify participants with normal cognition from participants with AD. Subsequently the model with best classification performance was used for predicting clinical outcome of MCI participants.

ii. Partial least square discriminant analysis model

Model description: Partial least squares regression (PLS regression) is a statistical method that bears some relation to principal components regression; instead of finding hyperplanes of maximum variance between the response and independent variables, it finds a linear regression model by projecting the predicted variables and the observable variables to a new space. Because both the *X* and *Y* data are projected to new spaces, the PLS family of methods are known as bilinear factor models.
Title of the study: An exploration of mortality risk factors in non-severe pneumonia in children using clinical data from Kenya
Link to the study: https://doi.org/10.1186/s12916-017-0963-9
Domain: Medical care
Study design: Retrospective cohort study
Data sources used: Hospital medical admission and discharge reports and laboratory data
Use of model to determine: To identify factors that best discriminate inpatient mortality risk in non-severe pneumonia and explore whether these factors offer any added benefit over the current criteria used to identify children with pneumonia requiring inpatient care.
Type of model:
Models and parameters used in the study: In this study, the machine learning models used in analysis were partial least squares - discriminant analysis (PLS-DA), random forests (RFs), support vector machines (SVMs) and elastic nets. Model validation as checked by employing a 10-fold internal cross

validation on two thirds of the data. The remaining one third of the data was used as the validation set. The selection of critical parameters for each of these modelling techniques was auto-determined by the R caret train function by choosing the tuning parameters that produced the highest values of receiver operating characteristic (ROC) curves where a grid search crossvalidation was applied.

iii. Decision tree learning

Model description: Decision tree learning is one of the predictive modelling approaches used in statistics, data mining and machine learning. It uses a decision tree (as a predictive model) to go from observations about an item (represented in the branches) to conclusions about the item's target value (represented in the leaves). Tree models where the target variable can take a discrete set of values are called classification trees; in these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels. Decision trees where the target variable can take continuous values (typically real numbers) are called regression trees. Decision trees are among the most popular machine learning algorithms given their intelligibility and simplicity.
Title of the study: Application of Machine Learning to Predict Dietary Lapses During Weight Loss
Link to the study: https://doi.org/10.1177/1932296818775757
Domain: Health status monitoring
Study design: An online survey
Data sources used: An online Weight Watchers program (i.e., an evidence-based program) to loss the weight
Use of model to determine: prediction of dietary lapses during weight loss
Type of model: Predictive model
Models and parameters used in the study: In this study, the optimal group model was identified using ensemble methods (e.g., combining weighted vote of predictions from Random Forest, Logit. Boost, Bagging, Random Subspace, Bayes Net). Cost-sensitive methods were used by incorporating a cost matrix (e.g., a matrix of penalties for misclassification) into each decision tree. Cost sensitive penalties were selected based on a balance of sensitivity and specificity (e.g., highest possible sensitivity while maintaining adequate specificity).

iv. Random forest

Model description: Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees. Random decision forests correct for decision trees habit of overfitting to their training set.
Title of the study: Machine learning models in breast cancer survival prediction
Link to the study: https://pubmed.ncbi.nlm.nih.gov/26409558/
Domain: Medical care
Study design: Cohort study
Data sources used: A dataset with eight attributes that include the records of 900 patients in which 876 patients (97.3%) and 24 (2.7%) patients were females and males respectively
Use of model to determine: To propose a rule-based classification method with machine learning techniques for the prediction of different types of Breast cancer survival.
Type of model: Prediction model
Models and parameters used in the study: In this study, following models were used for the prediction of breast cancer survival Naive Bayes (NB), Trees Random Forest (TRF), 1-Nearest Neighbor (1NN), AdaBoost (AD), Support Vector Machine (SVM), RBF Network (RBFN), and Multilayer Perceptron (MLP) machine learning techniques with 10-cross fold technique. The performance of machine learning techniques were evaluated with accuracy, precision, sensitivity, specificity, and area under ROC curve.

v. Gradient Boosting Classifier (GBC)

Model description: Gradient boosting is a machine learning technique for regression and classification problems, which produces a prediction model in the form of an ensemble of weak prediction models, typically decision trees. It builds the model in a stage-wise fashion like other boosting methods do, and it generalizes them by allowing optimization of an arbitrary differentiable loss function.
Title of the study: Machine learning for characterizing risk of type 2 diabetes mellitus in a rural Chinese population: the Henan Rural Cohort Study
Link to the study: https://doi.org/10.1038/s41598-020-61123-x
Domain: Health status monitoring
Study design: Cohort study
Data sources used: Data on socio-demographic characteristics, information on physical examination, and laboratory test data
Use of model to determine: To test the ability of machine learning algorithms for predicting risk of type 2 diabetes mellitus (T2DM)
Type of model: Predictive model
Models and parameters used in the study: In this study, risk assessment models for T2DM were developed using six machine learning algorithms, including logistic regression (LR), classification and regression tree (CART), artificial neural networks (ANN), support vector machine (SVM), random forest (RF) and gradient boosting machine (GBM). The model performance was measured in an area under the receiver operating characteristic curve, sensitivity, specificity, positive predictive value, negative predictive value and area under precision recall curve. The importance of variables was identified based on each classifier and the shapley additive explanations approach. Using all available variables, all models for predicting risk of T2DM demonstrated strong predictive performance, with AUCs ranging between 0.811 and 0.872 using laboratory data and from 0.767 to 0.817 without laboratory data. Among them, the GBM model performed best (AUC: 0.872 with laboratory data and 0.817 without laboratory data). Performance of models plateaued when introduced 30 variables to each model except CART model. Among the top-10 variables across all methods were sweet flavor, urine glucose, age, heart rate, creatinine, waist circumference, uric acid, pulse pressure, insulin, and hypertension. New important risk factors (urinary indicators, sweet flavor) were not found in previous risk prediction methods, but determined by machine learning in our study. Through the results, machine learning methods showed competence in predicting risk of T2DM, leading to greater insights on disease risk factors with no priori assumption of causality.

vi. k-nearest neighbours/k-means

Model description: In pattern recognition, the *k*-nearest neighbors algorithm (*k*-NN) is a non-parametric method proposed by Thomas Cover used for classification and regression. In both cases, the input consists of the *k* closest training examples in the feature space. The output depends on whether *k*-NN is used for classification or regression:

- In *k-NN classification*, the output is a class membership. An object is classified by a plurality vote of its neighbors, with the object being assigned to the class most common among its *k* nearest neighbors (*k* is a positive integer, typically small). If *k* = 1, then the object is simply assigned to the class of that single nearest neighbor.
- In *k-NN regression*, the output is the property value for the object. This value is the average of the values of *k* nearest neighbors.

Title of the study: Machine Learning With K-Means Dimensional Reduction for Predicting Survival Outcomes in Patients With Breast Cancer
Link to the study: https://doi.org/10.1177/1176935118810215
Domain: Health care
Study design: Prospective cohort study
Data sources used: Clinicopathological and genomic data
Use of model to determine: to integrate multiple clinicopathological and genomic factors with dimensional reduction across machine learning algorithms to compare survival predictions.
Models and parameters used in the study: In this study, to predict survival outcome, a total of 27 features (including indicator variables) from the 18 clinicopathological features mentioned above and 1 genomic feature were used to construct the models. We trained a series of nonlinear machine learning methods with 10-fold cross-validation of the training set upon 10 random training/validation splits using Gradient Boosting (R package *xgboost*), Random Forest (R package *random Forest*), SVM with a radial basis (SVM, R package *svm*), and ANN (R package *nnet*). The 10 random training/validation splits included the same patients in each set as those for K-means clustering above. For each split, 80% of the analytic cohort were randomly selected as our training dataset. Model performance was examined in the remaining 20% validation dataset, by estimating ROC, accuracy, and CS.

vii. Support Vector Machine

Model description: In machine learning, support-vector machines (SVMs, also support-vector networks) are supervised learning models with associated learning algorithms that analyze data used for classification and regression analysis.
Title of the study: A machine learning-based framework to identify type 2 diabetes through electronic health records
Link to the study: https://doi.org/10.1016/j.ijmedinf.2016.09.014
Domain: Health status monitoring
Study design: Cohort study
Data sources used: Electronic health records
Use of model to determine: to develop a semi-automated framework based on machine learning as a pilot study to liberalize filtering criteria to improve recall rate with a keeping of low false positive rate
Models and parameters used in the study: In this study, several widely-used classification model such as k-Nearest-Neighbors (kNN), Naïve Bayes (NB), Decision Tree (J48), Random Forest (RF), Support Vector Machine (SVM) and Logistic Regression (LR) to model patterns of cases and controls based on our extracted features and then use the models to test the ability of our extracted features on identifications of T2DM subjects.

viii. Neural networks

Model description: These are the systems modeled after the human brain, mimicking the ways we learn and make decisions. These networks consist of input and output layers, as well as hidden layers, similar to the neural networks in our brains.
Title of the study: Machine learning approaches to the social determinants of health in the health and retirement study
Link to the study: https://doi.org/10.1016/j.ssmph.2017.11.008
Domain: social determinants
Study design: Retrospective cohort
Data sources used: Health and retirement study database
Use of model to determine: To investigate how machine learning may add to our understanding of social determinants of health using data from the Health and Retirement Study.

Models and parameters used in the study: To assess different machine learning methods' ability to predict the biomarkers of interest, we first considered two OLS (Ordinary Least Square) regression models. The first was minimal and included gender, age, and age squared. The second was based on current understanding of social determinants of health, particularly that education and economic position have demonstrated associations with health. This theory-based model was parsimonious and included, as linear variables, household income, household wealth, and two binary variables indicating a high school-level education and less than a high school-level education, in addition to the parameters in the minimal model.
We next consider four machine learning algorithms: repeated linear regressions - akin to genome-wide association studies (GWAS), penalized linear regressions (Hastie, 2009), random forests (Breiman, 2001), and neural networks (Kriesel, 2007). These cover parametric and nonparametic approaches, with varying abilities to account for nonlinearity. While it is not possible to consider all machine learning algorithms, in addition to the broad coverage offered by these algorithms, all have been used in the medical literature (Patel et al., 2010, Rehkopf and Laraia, 2011, Horvath, 2013, Kapetanovic et al., 2004, Sato et al., 2005, Goldstein et al., 2010) and penalized regressions and random forests are particularly commonly-taught methods (Hastie et al., 2009, Bishop, 2006). These four also offer some prospect for interpretation rather than being completely "black box" approaches.

ix. Hierarchical clustering

Model description: In data mining and statistics, hierarchical clustering (also called hierarchical cluster analysis or HCA) is a method of cluster analysis which seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:[1]

- Agglomerative: This is a "bottom-up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
- Divisive: This is a "top-down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Title of the study: Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables
Link to the study: https://doi.org/10.1016/S2213-8587(18)30051-2
Domain: Health status monitoring
Study design: A data driven cluster analysis
Data sources used: Diabetes registry
Use of model to determine:
Models and parameters used in the study: In this study, a data-driven cluster analysis (k-means and hierarchical clustering) in patients with newly diagnosed diabetes (n=8980) from the Swedish All New Diabetics in Scania cohort. Clusters were based on six variables (glutamate decarboxylase antibodies, age at diagnosis, BMI, HbA1c, and homoeostatic model assessment 2 estimates of ß-cell function and insulin resistance), and were related to prospective data from patient records on development of complications and prescription of medication.

x. XGBoost

Model description: XGBoost is an open-source software library that provides a machine learning method of regression and classification using ensemble learning with gradient tree boosting (GTB). This software provides a gradient boosting framework for C++, Java, Python, R, Julia, Perl, and Scala. It works on Linux, Windows, and macOS. From the project description, it aims to provide a "Scalable, Portable and Distributed Gradient Boosting (GBM,

GBRT, GBDT) Library". It runs on a single machine, as well as the distributed processing frameworks Apache Hadoop, Apache Spark, and Apache Flink.

Title of the study: Prediction of Glucose Metabolism Disorder Risk Using a Machine Learning Algorithm: Pilot Study

Link to the study: https://pubmed.ncbi.nlm.nih.gov/30478026/

Domain: Health status monitoring

Study design: Retrospective cohort study

Data sources used: Medical records

Use of model to determine: To predict the risk of developing diabetes or GMD (Glucose Metabolism Disorder) using data from thousands of OGTTs (Oral Glucose Tolerance Test) and a machine learning technique (XGBoost)

Models and parameters used in the study: XGBoost is open-source software that provides a machine learning method of regression and classification using ensemble learning with gradient tree boosting (GTB). For each study, to apply supervised machine learning methods, the required label data was prepared. If a subject was diagnosed with diabetes or GMD at least once during the period, then that subject's data obtained in previous trials were classified into the risk group (y=1). After data processing, 13,581 and 6760 OGTTs were analyzed for study 1 and study 2, respectively. For each study, a randomly chosen subset representing 80% of the data was used for training 9 classification models and the remaining 20% was used for evaluating the models. Three classification models, A to C, used XGBoost with various input variables, some including OGTT data. The other 6 classification models, D to I, used LR for comparison.

Sciensano | Rue Juliette Wytsmanstraat 14 |
1050 Brussels | Belgium| e-mail: infact.coordination@sciensano.be|
Website: www.inf-act.eu |Twitter: @JA_InfAct