Haneef et al. Archives of Public Health (2022) 80:9
https://doi.org/10.1186/s13690-021-00770-6

Archives of Public Health

METHODOLOGY Open Access

Methodological guidelines to estimate population-based health indicators using linked data and/or machine learning techniques

Romana Haneef[1*], Mariken Tijhuis[2], Rodolphe Thiébaut[3,4,5], Ondřej Májek[6,7], Ivan Pristaš[8], Hanna Tolenan[9] and Anne Gallay[1]

Abstract

Background: The capacity to use data linkage and artificial intelligence to estimate and predict health indicators varies across European countries. 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 resulting in reduced interoperability at various levels and timeliness, 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 calculating population-based health indicators to guide European countries using linked data and/or machine learning (ML) techniques with new methods.

Method: We have performed the following step-wise approach systematically to develop the methodological guidelines: i. Scientific literature review, ii. Identification of inspiring examples from European countries, and iii. Developing the checklist of guidelines contents.

Results: We have developed the methodological guidelines, which provide a systematic approach for studies using linked data and/or ML-techniques to produce population-based health indicators. These guidelines include a detailed checklist of the following items: rationale and objective of the study (i.e., research question), study design, linked data sources, study population/sample size, study outcomes, data preparation, data analysis (i.e., statistical techniques, sensitivity analysis and potential issues during data analysis) and study limitations.

Conclusions: This is the first study to develop the methodological guidelines for studies focused on population health using linked data and/or 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 methodologies using more linked data and ML-techniques to develop a structured cross-disciplinary approach for improving the population health information and thereby the population health.

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

* Correspondence: Romana.HANEEF@santepubliquefrance.fr
[1]Department of Non-Communicable Diseases and Injuries, Santé Publique France, Saint-Maurice, France
Full list of author information is available at the end of the article

Background

The availability of data generated from different sources is increasing as well as the possibility to link these data sources with other databases. More efficient ways of data linkage and the use of artificial intelligence (i.e., machine learning techniques) are required to generate comparable and timely health information across European countries. Using these innovative techniques has several advantages such as data linkage improving completeness and comprehensiveness of information to guide health policy processes [1]. New approaches more or less based on artificial intelligence allow us to handle data with a large number of dimensions (features) and units (feature vectors) more efficiently and with high precision. Many countries have already invested in the linkage including both deterministic and probabilistic linkages and linking their traditional health administrative data with other types of data and has increased interoperability [2]. The capacity to use data linkage and artificial intelligence (AI) to estimate and predict health indicators varies across European countries [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, interoperability issues (legal, organizational, semantic and technical levels), availability of a large number of variables, lack of skills and capacity to link and analyze big data [4]. Due to varying health information systems across European countries, makes challenging to learn from each other experiences.

To our knowledge, there are no methodological guidelines available, which could systematically guide countries in using linked data and/or machine learning techniques (ML-techniques) to estimate health indicators for population health research and monitoring. 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/or ML-techniques with new methods/techniques. InfAct (Information for Action) project is a joint action of Member States (MSs) aiming to develop a more sustainable EU (European Union) - health information system through improving the availability of comparable, robust and policy-relevant health status data and health system performance information [5]. InfAct gathers 40 national health authorities from 28 MSs.

The main objective of this study was to develop the methodological guidelines to estimate population-based health indicators using linked data and/or ML-techniques with new methods.

Methodology

We have performed following step-wise approach systematically to develop the methodological guidelines: i. scientific literature review, ii. identification of inspiring examples from European countries and iii. Developing the checklist of guidelines contents.

Literature review

Firstly, we performed a literature search to identify published articles focusing on estimating health indicators using linked data and/or machine learning techniques in the field of health surveillance and health care performance on August 1, 2020. We did not specify any time period to search for the related published articles as to obtain a wide range of studies published at any time. We included in our search peer-reviewed methodological articles, 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. We defined two search strategies to extract the citations from Pubmed with following keywords: search strategy 1 (Linked data, Machine learning techniques and Guidelines) and search strategy 2 (Health indicators, Linked data, Machine learning techniques and Guidelines). Further details of search strategies are reported in additional file 1. Based on this literature review, we identified various methodological approaches using linked data and/or machine learning techniques to develop these guidelines.

Identifying inspiring examples

We defined inspiring examples as those studies that take into account the use of linked data and/or ML-techniques to estimate health indicators and implied the estimated 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. We asked ten European countries who were part of InfAct project and have been performed studies using linked data and/or machine learning techniques [6].

Developing the checklist of methodological guidelines contents

Using the results of first two steps, we reviewed the method section of selected studies and have developed a checklist including the following 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 preparation, data analysis and study limitations.

Table 1 Methodological guidelines using linked data and/or machine learning techniques to estimate population-based indicators, a study performed under InfAct project, May 2021

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	State the significance level of alpha and power based on the defined research question to calculate the sample size.	☐
5	**Study outcomes**		
5.1	Main outcomes	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	Describe the level of estimation of health outcomes at the lowest possible granularity level (i.e., at community, metropolitan, departmental or regional levels).	☐
6	**Data preparation**		
6.1	A. Data extraction	Extract data with required input variables from linked data set to a single file or a spreadsheet that could be converted according to the required format of the statistical software for data analysis.	☐
6.2	Coding of variables	Code the input variables, which are common in different linked data sets continuous or categorical or binary variables for required data analysis.	☐
	B. Data preparation to develop and apply a ML-algorithm		
6.3		Identify and define the target groups for a given defined time window based on the outcome of interest.	☐
6.4		Code the inputs variables, which are common in different linked data sets to continuous or categorical or binary variables for a given defined time window time.	☐
6.4		Split of final data set into 80% training and 20% test data set.	☐
7	**Data analysis**		
7.1	A. Variables selection	Select variables after the removal of all variables with a variance equal to zero.	☐
7.2		Estimate the RelifExp score based on the relevance of each variable to the outcome of interest.	☐
	B. Statistical techniques		
7.3	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		
7.4		Train various models and compare the performances of each model in terms of AUC curve (only for binary classifier).	☐
7.5		Validate the model performance using k-fold cross-validation first on training data set, and then assess the model performance on test data set.	☐
7.6		Select the final model based on specific performance metrics including sensitivity, specificity, PPV*, NPV*, F1-score and kappa.	☐
	C. Sensitivity/uncertainty analysis		
7.7		Perform a sensitivity analysis to identify the most influential parameters for a given output of a model.	☐
7.8		Select an appropriate method to perform the sensitivity analysis.	☐
7.9		Calculate the uncertainty in estimates using 95% CI* and describe the source of uncertainty (if relevant).	☐

Table 1 Methodological guidelines using linked data and/or machine learning techniques to estimate population-based indicators, a study performed under InfAct project, May 2021 *(Continued)*

Item number	Checklist item	Description	
	D. Potential issues during data analysis		
	I. Missing data		
7.10		Identify the missing data in the given dataset.	☐
7.11		Apply an appropriate technique for the imputation of missing values in the given data set.	☐
7.12	II. Imbalanced target group in a given dataset	Apply an appropriate technique to create a balanced data set either using down sampling or over sampling approach.	☐
7.13	III. Bias and variance tradeoff	Find the most generalizable model to keep the balance between bias and variance.	☐
8	**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

Expected outcomes

The methodological guidelines to estimate health indicators focused on population health research using linked data and/or ML-techniques.

Results

Literature review

We reviewed 215 citations from PubMed and 118 were included in our final sample to develop these methodological guidelines (Fig. 1). Sixteen additional studies (i.e., inspiring examples from European countries) identified from InfAct project were also included to the final sample. The final sample included 134 studies using linked data and/or machine learning techniques to address various research questions either to describe or estimate 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 [7, 8]. We reviewed the methodologies applied in the selected studies and developed a checklist of various steps that could be adopted systematically to calculate health indicators using linked data and/or ML-techniques.

Inspiring examples

We have identified 16 studies as inspiring examples from ten European countries. These studies adopted various methodological approaches to estimate health indictors, either by using data linkage (12 studies), machine learning methods (2 studies) or both data linkage and machine learning approaches (2 studies). These studies were used to develop these guidelines.

Methodological guidelines for studies using linked data and/or ML-techniques

We have developed a checklist of key methodological steps that are recommended to adopt systematically while calculating population-based health indicators (Table 1) and include the following items as methodological guidelines with examples of studies:

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 problem, I = Intervention/exposure/risk factor, C = Comparator/control/alternative intervention [if appropriate], O = Outcome) are used in evidence-based practice to frame and answer clinical and health care-related questions [9]. These criteria could be adopted according to population health research questions. The research questions should be simple and smart for example "for obese children (population), does the use of community recreation activities (intervention) compared to educational programs on lifestyle changes (comparator) reduce the risk of diabetes mellitus (outcome)? The population based studies commonly focused on 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.

Study design

The second step is to select the appropriate study design that could best address the proposed research question. The choice of study design may depend on the type of available data. 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 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 studies (for comparing exposure between cases and controls), etc.

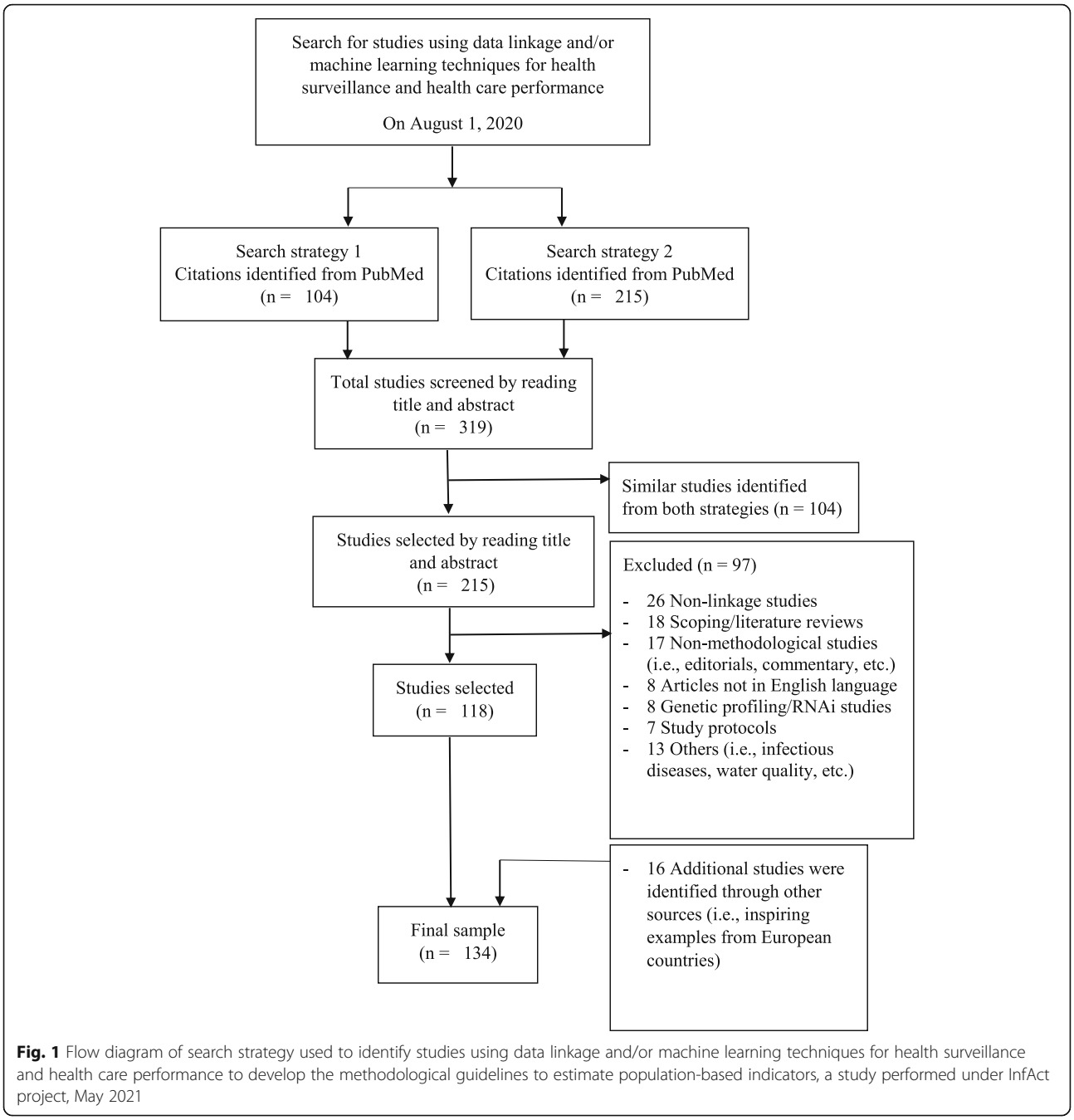

Fig. 1 Flow diagram of search strategy used to identify studies using data linkage and/or machine learning techniques for health surveillance and health care performance to develop the methodological guidelines to estimate population-based indicators, a study performed under InfAct project, May 2021

Linked data sources

The third step is to select the required linked data sources or one large data source to answer the proposed research question with specific objective. The linked data sources added the value and complement information on various factors which may have direct or indirect influence on health indicators. 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.

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 may allow having a large sample size and adds to large number of variables for analysis. The large sample size allows stratified analysis among sub-groups. The inclusion and exclusion criteria of the study population should be clearly defined according to the research question. The age, sex of included sample and the period of data collection should be clearly stated.

The null and alternative hypothesis should be clearly stated based on the research question [10]. The commonly used statistical significance values of alpha are 0.01, 0.05 or 0.1. To calculate the sample size, significance level of alpha and power should be stated based on the defined research question.

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 [9].

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 various geographical level and to adopt the health decisions according to the local needs.

Data preparation

The six step involve data preparation with two possibilities: raw data extraction and developing and applying ML-algorithm (Table 1).

Raw data extraction This step involves raw 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 data file or a spreadsheet that could be converted to different file formats according to the statistical software to be used for data analysis. The input variables, which are common in different linked data sets could be coded to binary or continuous or categorical variables. The dates are coded as a continuous variable. The linked data could add some redundancies while linking some variables to extract one specific information and to perform specific analyses.

To develop and apply a ML-algorithm This step involves the preparation of data to develop and apply an ML-algorithm, following sub-steps:

I. *Target/Case definition:* First, the targets are identified and defined based on the outcome of interest 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 are defined as positive targets and non-diabetes patients in last 6 months are defined as negative targets).
II. *Coding of variables for a given time window:* All the input variables, which are common in different linked data sources, are coded to binary or continuous or categorical variables for a given defined time window (e.g., either 6 or 12 months). The choice of a time window to code is important and should be selected based on the research question and study objectives.
III. *Split of final data into training and test data sets:* In most of the studies, 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 number of positive target - 1 group over the number of negative target - 0 group in the training dataset, a random down sampling or over sampling can be performed in the target 0 group to achieve the same number of individuals in both target groups. This helps to avoid the bias in ML-algorithm. Later, the selection of variables and the models is performed using the training data. The test data is used solely to test the final model performance.
According to the context of study and available data, different techniques of normalizing the data and cross-validation could be used.

Data analysis

The seventh step is the data analysis that may include variable selection, application of different statistical techniques, sensitivity/uncertainty analysis and some potential issues that may encounter during the data analysis (Table 1).

Variables selection First, all variables with a variance equal to zero are removed. Then the ReliefF exp. method could be applied (i.e., is a noise tolerant method and is not affected by features interactions) to estimate the score based on the relevance of each variable to the outcome of interest and to minimize the collinearity effect [11]. All variables are ranked according to the ReliefF exp. score and for continuous variables the score ranges from 0 to 1. For example, the cutoff score could be selected based on the visual inspection of the ordered plot of ReliefF values for all variables, called “elbow plot” approach (e.g., 0.01). In this case, the variables that had a

ReliefF exp. score equal or more than 0.01 could be included to train different models and the variables less than 0.01 could be excluded.

Statistical techniques There are several statistical techniques that are applied to linked data either using classical statistical techniques or with ML-techniques. The former may be used for regression and later for classification purposes. In general, both of these techniques could be used to estimate, classify and predict the population health indicators or to evaluate the health care interventions according to the available linked datasets. The brief description of different techniques is reported in additional file 2.

I. ***Classical statistical techniques:*** Several classical statistical techniques were identified in the selected studies to analyze the linked data set. Following are the most commonly used techniques: linear and logistic regression, Linear Discriminant Analysis (LDA) model [12, 13], multilevel linear regression [14], multivariate logistic regression [15], multivariable hierarchical modified Poisson regression [16], Cox regression models [17], LASSO regression [18, 19], Generalized Estimating Equation (GEE) models [20], inverse probability weighting methods [21], Blinder-Oaxaca decomposition method [22] and Markov modelling [23].
II. ***ML-techniques:*** Several ML-techniques are applied, which focused on health care research. These techniques could be adopted to population health studies. Following are the most commonly used supervised ML-techniques: linear and logistic regression, Linear Discriminant Analysis (LDA) model [12, 13], 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, following three main steps are used to train and select the final model:

a. *Training various models:* Some 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 receiver operating characteristic (AUC) curve. AUC curve is an evaluation metric for binary classification problems. It is a probability curve that plots TPR (true positive rates) or sensitivity against FPR (false positive rates) or I − Specificity at various threshold values and essential separates the 'signal' from the 'noise'. The AUC is the measure of the ability of a classifier to distinguish between classes [34]. The higher the AUC, the better the performance of the model at distinguishing between positive and negative classes.
b. *Model validation techniques:* To validate the model, k-fold cross-validation is 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 have been used as the testing set [35]. This technique allows to estimate the performance or accuracy of the model using data not utilized during training of the model.

After the first validation of the models using k-fold cross-validation on training data set, the model performances are assessed using the test data set.

iii. *Selection of final model*: After the model validation, the algorithm selection process is automated by giving the computer a 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.

Sensitivity/uncertainty analysis After the selection of final model, sensitivity analysis is performed. This analysis refers to identifying the most influential assumptions or 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.

Potential issues during data analysis During the data analysis, following are some common issues, which may

encounter: missing data, imbalanced datasets and bias-variance tradeoff.

I. ***Missing data:*** In datasets (small or big), 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 type of studies using classical statistical methods and ML-techniques:

a. ***For classical statistical methods:*** 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 (e.g. linear regression) for handling missing data (such as complete case analysis, overall mean/mode/median imputation, and the missing-indicator method) are more feasible to apply but may produce biased results. Multivariate Imputation by Chained Equation (MICE) is a multiple imputation techniques and does not avoid all bias but may be less prone to bias and does not help with MNAR [40, 41].
b. ***For ML-studies:*** There are eight most common ways to replace the missing values, which could be applied in both non-ML and ML-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]. These techniques are also robust to MAR data. 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:*** Second issue is the imbalanced dataset (i.e., the number of positive and negative targets/cases/values are unequal.) 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 down sampling and the second one is over sampling [45, 46].
III. ***Bias and variance tradeoff:*** 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 is also the rational for cross-validation approaches. This balance is key to finding the most generalizable model [47].

Model tuning/hyperparameter tuning: It is an important step to improve the model performance and accuracy. Robust model tuning provides insight on how model structure and hyperparameters influence the model performance [48]. Hyperparameters are adjustable parameters that must be tuned in order 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].

Study limitations

Study limitations are important and should be reported to addressing various issues for further research. Different studies using data linkage and/or 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 to compare the data, lack of standard methods for data collection or inability to link some cases due to incorrect identifier); ***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); ***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., appropriate choice of a time window to code the variables to estimate the incidence, overfit or underfit of the model used in ML-studies, boosted algorithm may require a high computational capacity).

Discussion

Main results

We have developed a checklist of eight items as the methodological guidelines, which provide a systematic approach using linked data and/or machine learning techniques to produce population-based health indicators.

There are few studies available that describe the reporting guidelines for linked data focused on population health research. For example, one study illustrates the guidelines to evaluate the methodological quality of studies using linked data and to report their results in a consistent manner [8]. Another 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 [7]. Another study developed TIDieR-PHP (Template for Intervention Description and Replication-Population Health and Policy) checklist to improve the reporting of PHP (population health and policy) interventions [50]. The STROBE (Strengthening the Reporting of Observational Studies in Epidemiology) guidelines were developed for reporting of observational studies [51]. The STRATOS (STRengthening analytical thinking for observational studies) initiative was taken to provide accessible and accurate guidance in the design and analysis of observational studies in medical research [52]. These guidelines could also be used for population health research studies. All these guidelines are important to improve the design, analysis and reporting of results. Nevertheless, the existing reporting guidelines are not fully designed to capture key methodological aspects applied to linked data and/or ML techniques for population health research.

Scope

These guidelines define a systematic approach for studies using linked data and/or ML-techniques to estimate health indicators for population health research. We used peer-reviewed published methodological studies, which applied data linkage and ML-techniques 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 of methods to be used for the calculation of health indicators and are flexible enough to integrate new methods used for population health research over time.

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. These guidelines would also allow to harmonize and practice certain methodological approaches to perform comparative studies between countries. 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 with a systematic approach to perform studies using data linkage and/or machine learning techniques to calculate health indicators for population health research. Moreover, these guidelines would improving the quality of research methods. We have provided at least one example of a study that has used the reported statistical techniques to better understand different aspects.

There are few limitations: *first,* we provided a systematic approach with general and basic techniques that are most commonly applied for studies using data linkage and/or 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. However, the reported techniques cover the main and basic techniques, which are commonly applied. *Second,* there are more studies possible, which have applied these techniques and are not reported in this study. Though the studies which are reported here, have covered the important techniques. *Third,* we did not perform a systematic review to identify the studies. Nevertheless, the adopted search strategy allowed to identify relevant studies and covered all the basic aspects of studies.

Recommendations

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

Data sources: data related to employment, education, occupation and socioeconomic status should be readily available/accessible to enrich the analyses 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 [53]. *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 technique is important to have results that are more robust. *Knowledge translation*: better approaches are required to communicate the estimates to the policymakers and other public stakeholders. This is key to evidence-informed policy-making and to support decision making about the allocation of resources. *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. Moreover, to develop joint country studies among European countries on using machine-learning techniques for public health research are needed.

Conclusions

This is the first study to develop the methodological guidelines to estimate population-based health indicators using linked data and/or 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 structured cross-disciplinary approach for improving the population health information and thereby the population health.

Abbreviations

EU: European Union; MSs: Member States; AI: Artificial Intelligence; ML-techniques: 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 Eq.; LS-SVM: Lease Square-Support Vector Machine; PPV: Positive Predictive Value; NPV: Negative Predictive Value; GDPR: General Data Protection Regulations

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s13690-021-00770-6.

Additional file 1. It describes the search strategies used to identify citations related to data linkage and/or machine learning technique used for studies focused on health status monitoring and health care.

Additional file 2. It describes the various statistical techniques used for data analysis using both classical statistical techniques and ML-techniques.

Authors' contributions

Conceived and designed the study: RH and AG. Performed the study: RH AG. Analyzed the data: RH. Developed the checklist of items RH AG: Writing the first draft of the guidelines: RH. Contributed to the writing of the manuscript: All authors contributed to the writing of the manuscript. All authors have read and approved the final manuscript.

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).

Availability of data and materials

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

All authors gave the consent for publication.

Competing interests

R. Haneef is the first author of this paper and the section editor of "health information system" of "Archives of Public Health". All other authors declare that they have no competing interests related to the work.

Author details

[1]Department of Non-Communicable Diseases and Injuries, Santé Publique France, Saint-Maurice, France. [2]National Institute for Public Health and the Environment (RIVM), Bilthoven, The Netherlands. [3]Bordeaux University, Bordeaux School of Public Health, Bordeaux, France. [4]INSERM / INRIA SISTM team, Bordeaux Population health, Bordeaux, France. [5]Medical Information Department, Bordeaux University Hospital, Bordeaux, France. [6]Institute of Health Information and Statistics of the Czech Republic, Prague, Czech Republic. [7]Institute of Biostatistics and Analyses, Faculty of Medicine, Masaryk University, Brno, Czech Republic. [8]National Institute of public health, division of health informatics and biostatistics, Zagreb, Croatia. [9]Finnish Institute for Health and Welfare (THL), Helsinki, Finland.

Received: 27 May 2021 Accepted: 17 December 2021
Published online: 04 January 2022

References

1. Lloyd K, McGregor J, John A, Craddock N, Walters JT, Linden D, 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. Schizophr Res. 2015;166(1):131–6. https://doi.org/10.1016/j.schres.2015.05.036.
2. Delnord M, Szamotulska K, Hindori-Mohangoo AD, Blondel B, Macfarlane AJ, Dattani N, et al. Linking databases on perinatal health: a review of the literature and current practices in Europe. Eur J Pub Health. 2016;26(3):422–30. https://doi.org/10.1093/eurpub/ckv231.
3. Haneef R, Delnord M, Vernay M, Bauchet E, Gaidelyte R, Van Oyen H, et al. Innovative use of data sources: a cross-sectional study of data linkage and artificial intelligence practices across European countries. Arc Public Health. 2020;78(1):55. https://doi.org/10.1186/s13690-020-00436-9.
4. Bradley CJ, Penberthy L, Devers KJ, Holden DJ. Health Services Research and Data Linkages: Issues, Methods, and Directions for the Future. *Health Serv Res*. 2010;**45**(5p2):1468–88.
5. Joint Action on Health Information: https://www.inf-act.eu/. 2018.
6. INFACT: Inspiring Examples from European Countries: https://www.inf-act.eu/sites/inf-act.eu/files/2021-02/D%209.2%20%28Part%20B%29_Inspiring%20Examples.pdf. 2020.
7. Stevens G, Alkema L, Black R, Boerma J, Collins G, Ezzati M, et al. Guidelines for accurate and transparent health estimates reporting: the GATHER statement. Lancet. 2016;388(10062):e19–23. https://doi.org/10.1016/S0140-6736(16)30388-9.
8. 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):486–9. https://doi.org/10.1111/j.1753-6405.2011.00741.x.
9. Illinois Uo: What is a PICO model?:https://researchguides.uic.edu/c.php?g=252338&p=3954402. 2020.
10. Brownlee J: A Gentle Introduction to Statistical Hypothesis: https://machinelearningmastery.com/statistical-hypothesis-tests/. 2018.
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. Ezzati A, Zammit AR, Harvey DJ, Habeck C, Hall CB, Lipton RB, et al. Optimizing machine learning Methods to improve predictive models of Alzheimer's disease. J Alzheimers Dis. 2019;71(3):1027–36. https://doi.org/10.3233/JAD-190262.
13. Yang T, Zhang L, Yi L, Feng H, Li S, Chen H, 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. https://doi.org/10.2196/15431.
14. 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):e24–33. https://doi.org/10.1016/S2468-2667(17)30212-8.
15. Sultan A, West J, Grainge M, Riley R, Tata L, Stephansson O, et al. Development and validation of risk prediction model for venous thromboembolism in postpartum women: multinational cohort study. Bmj. 2016;**5**(355). https://doi.org/10.1136/bmj.i6253.
16. 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):2041–51. https://doi.org/10.1161/CIRCULATIONAHA.117.030488.
17. 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. https://doi.org/10.1136/bmjopen-2018-026507.
18. Odgers D, Tellis N, Hall H, Dumontier M. Using LASSO regression to predict rheumatoid arthritis treatment efficacy. AMIA Jt Summits Transl Sci Proc. 2016;20:176–83.
19. Orriols L, Avalos-Fernandez M, Moore N, Philip P, Delorme B, Laumon B, et al. Long-term chronic diseases and crash responsibility: a record linkage study. Accid Anal Prev. 2014;71:137–43. https://doi.org/10.1016/j.aap.2014.05.001.
20. 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;2:841–9. https://doi.org/10.1016/j.ssmph.2016.10.016.
21. Astley CM, Chew DP, Keech W, Nicholls S, Beltrame J, Horsfall M, 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):475–82. https://doi.org/10.1016/j.hlc.2019.03.015.
22. 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(1):267. https://doi.org/10.1186/s12889-015-1607-7.
23. Asaria M, Walker S, Palmer S, Gale CP, Shah AD, Abrams KR, et al. Using electronic health records to predict costs and outcomes in stable coronary artery disease. Heart. 2016;102(10):755–62. https://doi.org/10.1136/heartjnl-2015-308850.
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. https://doi.org/10.1186/s12916-017-0963-9.
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):1045–52. https://doi.org/10.1177/1932296818775757.
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. https://doi.org/10.3233/THC-151071.
27. Rahimian F, Salimi-Khorshidi G, Payberah AH, Tran J, Ayala Solares R, Raimondi F, 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. https://doi.org/10.1371/journal.pmed.1002695.
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. Sci Rep. 2020;10(1):4406. https://doi.org/10.1038/s41598-020-61123-x.
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. https://doi.org/10.1177/1176935118810215.
30. Zheng T, Xie W, Xu L, He X, Zhang Y, You M, et al. A machine learning-based framework to identify type 2 diabetes through electronic health records. Int J Med Inform. 2017;97:120–7. https://doi.org/10.1016/j.ijmedinf.2016.09.014.
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;4:95–9. https://doi.org/10.1016/j.ssmph.2017.11.008.
32. Ahlqvist E, Storm P, Käräjämäki A, Martinell M, Dorkhan M, Carlsson A, et al. Novel subgroups of adult-onset diabetes and their association with outcomes: a data-driven cluster analysis of six variables. Lancet Diabetes Endocrinol. 2018;6(5):361–9. https://doi.org/10.1016/S2213-8587(18)30051-2.
33. Maeta K, Nishiyama Y, Fujibayashi K, Gunji T, Sasabe N, Iijima K, et al. Prediction of glucose metabolism disorder risk using a machine learning algorithm: pilot study. JMIR Diabetes. 2018;3(4):10212. https://doi.org/10.2196/10212.
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. edn.: 1–51.
37. Sensitivity Analysis: From Theory to Practice. In: Global Sensitivity Analysis The Primer. edn.: 237–275.
38. Variance-Based Methods. In: Global Sensitivity Analysis The Primer. edn.: 155–182.
39. Elementary Effects Method. In: Global Sensitivity Analysis The Primer. edn.: 109–154.
40. Donders ART, van der Heijden GJMG, Stijnen T, Moons KGM. Review: a gentle introduction to imputation of missing values. J Clin Epidemiol. 2006;59(10):1087–91. https://doi.org/10.1016/j.jclinepi.2006.01.014.
41. Chinomona A, Mwambi H. Multiple imputation for non-response when estimating HIV prevalence using survey data. BMC Public Health. 2015;15(1):1059. https://doi.org/10.1186/s12889-015-2390-1.

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):579–87. https://doi.org/10.1109/JBHI.2016.2634587.
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. https://doi.org/10.1177/2374289519873088.
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. https://doi.org/10.1136/bmj.k1079.
51. von Elm E, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. The Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. PLoS Med. 2007;4(10):e296. https://doi.org/10.1371/journal.pmed.0040296.
52. Sauerbrei W, Abrahamowicz M, Altman DG, le Cessie S, Carpenter J. STRengthening analytical thinking for observational studies: the STRATOS initiative. Stat Med. 2014;33(30):5413–32. https://doi.org/10.1002/sim.6265.
53. 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.

Publisher's Note