

**DARE**  
**DIGITAL LIFELONG PREVENTION**  
CODE NO. PNC0000002

Spoke 2 Deliverable  
**S2.D1.1 Process and outcome  
indicators in Spoke 2**

This research is co-funded by the Ministry of University and Research  
within the Complementary National Plan PNC-I.1  
"Research initiatives for innovative technologies  
and pathways in the health and welfare sector"

D.D. 931 of 06/06/2022, PNC0000002 DARE - Digital Lifelong Prevention



Ministero  
dell'Università  
e della Ricerca



**PNC**  
Piano nazionale per gli investimenti  
complementari al PNRR  
Ministero dell'Università e della Ricerca

**DARE** | Digital  
Lifelong  
Prevention

## S2.D1.1 Process and outcome indicators in Spoke 2

Deliverable information	
Spoke number and title	Spoke 2 – Community-based Digital Primary Prevention
WP number and title	WP1 - Ecosystem Building and Infrastructures
Related task(s)	T1.1
Lead beneficiary	UNIPA
Contributing beneficiaries	UNIPA, UKE, ARPA
Dissemination level	Public, fully open
Due date	15 June 2023
Actual date of delivery	14 June 2023
Author(s)	Vincenzo Conti (UKE), Francesco Prinzi (UNIPA), Giovanni Cicceri (UNIPA), Giosuè Lo Bosco (UNIPA), Salvatore Lo Verso (ARPA), Walter Mazzucco (UNIPA), Salvatore Vitabile (UNIPA)
Contributors	
Quality Assurance	Dario Gregori (UNIPD), Stefania Boccia (UCSC)

## Document history

Version	Date	Author(s) /Reviewer(s) (Beneficiary)	Description
0.1	2 May 2023	Vincenzo Conti (UKE), Francesco Prinzi (UNIPA), Giovanni Cicceri (UNIPA), Giosuè Lo Bosco (UNIPA), Salvatore Lo Verso (ARPA), Salvatore Vitabile (UNIPA)	First draft
0.2	25 May 2023	Vincenzo Conti (UKE), Francesco Prinzi (UNIPA), Giovanni Cicceri (UNIPA), Giosuè Lo Bosco (UNIPA), Salvatore Lo Verso (ARPA), Walter Mazzucco (UNIPA), Salvatore Vitabile (UNIPA)	Revision
0.3	12 June 2023	Task 1.1, Task 1.2, Task 1.3, Task 1.4	Final document
1.0	12 June 2023	HUB final check	Final version submitted

## Disclaimer

This publication reflects only the author's views, and the Funding Agency is not liable for any use that may be made of the information contained therein.



# Table of contents

Summary.....	5
List of abbreviations.....	6
1. Introduction.....	6
1.1. Objectives of the deliverable.....	6
1.2. Why are indicators important in DARE?.....	7
1.3. Which indicators? Clinical studies vs predictive model studies.....	7
1.4. List of indicators.....	8
2. The importance of data management in DARE.....	9
3. Indicators focused on Community-based Preventive Study evaluations.....	10
3.1. Detailed explanation of each indicator.....	10
4. Indicators focused on Machine Learning model evaluations.....	11
4.1. Detailed explanation of each indicator.....	11
5. Applicative Scenarios.....	15
6. Case Study Example: cardiovascular disease risk prediction as a primary prevention strategy.....	17
7. References.....	20



## Summary

The objective of the DARE project is to deliver innovative technologies that are effective, efficient, and readily deployable, thereby enabling the Italian Ministry of Health and the NHS to expedite the translational process in support of the prevention healthcare landscape.

Indicators are distinct markers that can be observed and measured, serving as evidence of progress toward achieving specific outputs or outcomes outlined in a logic model or work plan. This deliverable aims to appoint all partners of the DARE consortium with a comprehensive list of potential indicators. These indicators will be utilized to evaluate the usefulness and feasibility of implementing innovative preventive technologies during the pilot development phase.

Evaluating the impact of health problems, socioeconomic variables, geographical factors, and existing interventions is crucial for designing effective solutions. In addition, Machine Learning (ML) models have the potential to provide valuable insights and support decision-making in various real-world applications, developing at the same time a common framework for an integrated public health, environmental, and climate governance in support of the National Prevention Hub and the National System on Health, Environment, and Climate Protection. However, their actual utilization is hindered by the lack of interoperability. Evaluating ML models requires assessing their performance through multiple metrics based on specific tasks, including classification, regression, and segmentation. The richness of metadata associated with ML models is crucial for their interpretability, fairness, compliance with data regulations, and overall performance. The accuracy of correlating exposures with human health outcomes is vital for informed decision-making and effective interventions. Furthermore, ML model robustness, scalability, and interoperability with existing health systems enhance their reliability and practicality.

Therefore, this document aims to identify the essential indicators for evaluating the works conducted in SPOKE 2. The identified indicators serve the purpose of assessing activities that encompass significant community-based insights and also emphasize the real-world impacts of artificial intelligence methods for primary prevention.

## List of abbreviations

AI: Artificial Intelligence

AUC-ROC: Area Under the Receiver Operating Characteristic Curve

AWARE: Assess, WArn & Response

BMI: Body Mass Index

CDC: Center for Disease Control and Prevention

CML: Cascade Machine Learning

CVD: CardioVascular Disease

DL: Deep Learning

FAIR: Findable, Accessible, Interoperable, and Reusable

GRS: Global Risk Score

MAE: Mean Absolute Error

MDRO: Multi-Drug Resistant Organisms

MSE: Mean Squared Error

ML: Machine Learning

MLP: MultiLayer Perceptron

NHS: National Health Service

TB: Tuberculosis

VPDs: Vaccine-Preventable Diseases

## 1. Introduction

### 1.1. Objectives of the deliverable

The primary objective of this deliverable is to create a comprehensive framework of processes and outcome indicators that can be used to assess the effectiveness and feasibility of implementing digital preventive primary interventions. These indicators will serve as tools to measure the utility and implementability of these interventions.

## 1.2. Why are indicators important in DARE?

According to the Center for Disease Control and Prevention (CDC), an indicator is a "marker" that signifies achievement or progress. It refers to a specific, observable, and measurable accomplishment or change that demonstrates progress towards attaining a particular output or outcome, as outlined in a logic model or work plan [1].

Indicators play a crucial role in guiding the project workflow. They not only ease the evaluation process but also provide guidance on the direction it should undertake and the parameters needed to be considered. By defining specific indicators, the evaluation process can focus on collecting relevant data that aligns with the desired outcomes and objectives. This definition ensures that the evaluation process is targeted, efficient, and informed, allowing for meaningful analysis and decision-making. [2].

Indicators contribute to categorize the uncertainty that innovative technologies bring, thus ensuring that all relevant aspects are taken into account when the utility and implementability of innovative technologies have to be assessed from a public health perspective.

By streamlining the necessary stages throughout the pilot's development and execution, we will achieve a thorough evaluation that provides accurate responses and straightforward guidance. This document serves as a reference for researchers, facilitating the gathering, organizing, and sharing of essential metrics and results during the pilot's progression. Its purpose is to encourage the collection of pertinent utility and implementability indicators through data analysis and synthesis throughout the pilot's duration.

## 1.3. Which indicators? Clinical studies vs predictive model studies

DARE is a consortium among universities, research centers, research hospitals, private companies, and non-profit organizations, promoting the implementation of a wide variety of studies.

This SPOKE will focus on primary prevention through research and the development of dedicated pilots.

The tasks included in SPOKE 2 have a translational nature, encompassing traditional clinical studies fueled by digital technologies, as well as retrospective and prospective studies utilizing artificial intelligence (AI), machine learning (ML), and deep learning (DL) techniques. Consequently, the

identified indicators aim to assess both preventive studies (descriptive, observational, community trials) and associated predictive models.

### 1.4. List of indicators

The identified indicators can be categorized into three distinct groups. The first category comprises indicators designed to assess purely preventive studies (observational real-world studies, community trials). The second category focuses on evaluating specific aspects of Community-based Preventive Study. Lastly, some indicators are fundamental to both domains, serving as essential measures across both preventive studies and AI models-based studies. For this reason, the General indicators will be discussed both in Sections “Indicators focused on Community-based Preventive Study evaluations” and “Indicators focused on Machine Learning model evaluations”. Subsequently, Public health indicators will be discussed only for Community-based Preventive Study (Section 3.1), and ML model indicators for Machine Learning model evaluations (Section 4.1).

General indicators:

- Prospective Utility;
- Impact on Community Health;

Public health indicators:

- Economic Impact;
- Geographical impact;
- Existing Interventions;

ML model indicators:

- ML model evaluation;
- Metadata richness;
- Environmental-Health Correlation;
- ML model robustness;
- ML model scalability;
- Health ML systems interoperability;
- Data privacy and Security in ML solutions;
- Model Reliability;

- Reproducibility.
- Usability.

## 2. The importance of data management in DARE

Developing community-based primary preventive solutions relies heavily on data mining and data analytics techniques. However, the reliability of data-driven algorithms is being questioned due to their dependence on the quantity and, more importantly, the quality of the training data. Therefore, the establishment of a shared data-collection infrastructure, which is accessible and secured through specialized hardware, has become an indispensable resource for the effective implementation of community-based primary prevention in real-world practice.

Moreover, despite various disciplines and practices now involving the collection of data from healthcare sources (biological data, genomic data, and clinical data), public health surveillance systems, environment monitoring systems, health determinants data, and administrative data sources, there is a lack for data organization and standardization. This lack hinders the accessibility and usability of these data for the research community, resulting in increased difficulty and slower progress in research. By employing an online, shared, available, and easily accessible infrastructure, the work of researchers and other stakeholders can be facilitated and accelerated through agile access to diverse data and resources.

Building upon the foundation laid by the FAIR data principles established in 2016, one of the main objectives of the DARE project is to create an advanced data collection and processing infrastructure for SPOKE 2 participants and stakeholders. Following the principles of Findable, Accessible, Interoperable, and Reusable (FAIR) [3] data, this infrastructure aims at integrating and harmonizing different data sources, such as health, occupational, environmental, and climate data, in terms of their geographical, geometrical and semantic features and differences. This harmonization process is crucial in presenting a unified and conflict-free perception to users. A set of services will be developed and delivered for data access and processing to highlight any possible or unexpected correlation or association.

For these reasons, the developed activities will have to consider the needs above to guarantee the characteristics necessary for the actual use of AI systems.

### 3. Indicators focused on Community-based Preventive Study evaluations

In order to effectively evaluate the seriousness of a problem in community-based preventive study, it is essential to undertake a comprehensive assessment encompassing various dimensions. These dimensions include evaluating the impact on health, considering the burden experienced by communities and specific target groups, assessing the implications for public health, analysing economic considerations, and examining existing interventions. This multidimensional assessment serves as a valuable tool for researchers, epidemiologists, public health operators, and policymakers, enabling them to prioritize and address the most critical and significant issues within public health research and healthcare.

#### 3.1. Detailed explanation of each indicator

**Prospective Utility:** The implementation and training of machine learning models are designed to draw insights from new data, enabling the utilization of these trained models for planning and decision support. This encompasses various real-world applications, such as devising integrated public health, environmental, and climate strategies having substantial implications for community health. However, despite the great promise of many applications, their actual utilization is hindered by the absence of observational real-world studies and community trials [16]. This indicator aims to assess whether the system has been designed and/or evaluated through the abovementioned studies.

**Impact on Community Health:** Assessing the severity of the problem in terms of its impact on the health of individuals is crucial. This can include evaluating potential adverse health outcomes, such as morbidity or mortality rates associated with the condition or disease being studied.

**Economic Impact:** Evaluating the economic impact of the application is also relevant. This can involve assessing the costs associated with prevention, diagnosis, treatment, and management of the condition, as well as the potential for productivity losses or healthcare resource utilization.

**Geographical impact:** Geographical relevance of incidence factors influencing public health is essential for designing a prevention strategy. This includes territorial factors evaluation that can affect different models of disease diffusion and the relative impact on local health systems.

**Existing Interventions:** Considering the availability and effectiveness of existing interventions for the specific problem is essential. Evaluating the seriousness of the situation involves comparing the potential benefits of new interventions or approaches against the limitations or shortcomings of current standard-of-prevention options.

## 4. Indicators focused on Machine Learning model evaluations

A predictive model indicator refers to a measurement or a group of measurements used to evaluate how effectively a model can accurately predict specific outcomes. According to Kuhn and Johnson [4], selecting the right indicators is crucial for assessing the performance of predictive models. These indicators are essential in ensuring that the model can successfully achieve its intended tasks and goals. Without these indicators, it becomes difficult to determine whether the model functions properly or if improvements are required. The choice of appropriate indicators depends on the type of predictive model and the particular domain it is applied to, intending to provide valuable insights into the model's performance.

A multitude of factors can affect the power of a model to generalize effectively. These factors include but are not limited to the quantity and quality of training data available, the complexity of the model itself, and the protocols employed for evaluation and final validation. Each of these steps plays a crucial role in establishing and assessing the models, ensuring their practical applicability in real-world scenarios. Inadequate training data can lead to incomplete or biased learning, hindering the model's ability to make accurate predictions or classifications in real-world situations. An overly simplistic model may not capture the intricacies and nuances present in real-world data, resulting in poor generalization. On the other hand, an excessively complex model may overfit the training data, performing well on the training set but failing to generalize to new, unseen data. In general, proper validation procedures help verify the model's performance on unseen data, providing confidence in its ability to generalize beyond the training set.

### 4.1. Detailed explanation of each indicator

Based on the type of problem, different metrics need to be utilized for evaluating the model.

**Prospective Utility:** The implementation and training of machine learning models are designed to draw insights from new data, enabling the utilization of these trained models for planning and decision support. This encompasses various real-world applications, such as devising integrated public health, environmental, and climate strategies, having substantial implications for public health. However, despite the great promise of many applications, their actual utilization is hindered by the absence of observational real-world studies and community trials [16]. This indicator aims to assess whether the system has been designed and/or evaluated through the abovementioned studies.

**Impact on Community Health:** Assessing the severity of the problem in terms of its impact on the health of individuals is crucial. This can include evaluating potential adverse health outcomes, such as morbidity or mortality rates associated with the condition or disease being studied.

**ML model evaluation:** Before deploying predictive models, evaluating their performance is a critical factor. Various metrics are utilized, depending on the specific task that the model intends to solve, such as classification, regression, segmentation, etc. Due to the diverse, desirable properties of a model, no single metric can encapsulate them all. Consequently, it is common practice to report multiple metrics to summarize a model's performance comprehensively [5]. Evidence provides a comprehensive exposition of the commonly used metrics for evaluating binary classifiers. Metrics such as accuracy and AUC-ROC can give a general overview of the models' generalization capability. In addition, metrics such as specificity and sensitivity provide estimates of prediction capabilities for each class and help to evaluate the training in class-unbalanced scenarios. These metrics can be extended to assess multi-class classifiers through one-vs-rest or one-vs-one strategies. A discussion on segmentation metrics is provided in [6]. The Dice similarity coefficient and the Jaccard index are common metrics to evaluate segmentation tasks. A correct interpretation of such metrics is mandatory also to detect class imbalance and statistical bias issues. Then, [7] offers an overview of regression metrics, e.g., mean squared error (MSE) and mean absolute error (MAE), providing valuable insights into this area of evaluation.

**Metadata richness:** The number of environment tag (metadata) managed by the ML model is an important indicator in various ML and data processing scenarios. This indicator refers to the count of metadata [8] or environment tags associated with the data used or processed by the ML algorithm. This indicator can be relevant and have implications for various aspects of the model's

performance, interpretability, fairness, and compliance with data regulations. Striking the right balance between metadata richness and practicality is crucial for developing effective and reliable ML-based systems.

**Environmental-Health Correlation:** This indicator measures the ML model's efficiency level in identifying and quantifying the correlation between environmental conditions and/or climate and human well-being or health outcomes [9]. In such a context, "correlation" refers to the statistical relationship or association between environmental (e.g., air quality, temperature, pollution levels, etc.) and/or climate (temperature, humidity, etc.) variables and health-related metrics (e.g., disease prevalence and incidence, mortality rates, health complaints, etc.). This indicator is valuable in various fields, such as public health, environmental and/or climate research, and policy, where understanding the relationships between environmental conditions and/or climate and human health is extremely important for making well-informed decisions and effective interventions.

**ML model robustness:** The ML robustness indicator refers to the ability of the model to maintain its performance and generalization capabilities under various conditions and settings [10]. A robust ML model should produce consistent and reliable results, even when faced with noisy, incomplete, or adversarial data. A high value for this indicator indicates that the ML model is more reliable and robust in real health scenarios, making it a valuable resource for various applications. Robust models will be less likely to produce unreliable results, improving their practical utility and effectiveness in solving complex medical problems.

**ML model scalability:** This indicator refers to the ability of a machine learning model to efficiently handle increasingly large and complex data sets and computational demands. It assesses the degree to which the performance and training time of the model scale with respect to growth in data and resources [11]. A high value for this indicator indicates that the ML model can efficiently handle increasing volumes of data and computational requirements, making it a powerful and adaptable solution for data-intensive tasks. Scalable models are essential for handling big data scenarios and ensuring consistent performance in medical-world applications.

**Health ML systems interoperability:** This indicator refers to the seamless integration and interoperability of an ML solution with pre-existing health systems and infrastructure related to health management. It assesses the ML model's ability to integrate with and complement existing

systems, enhance their capabilities, and provide valuable information. A high value of this indicator demonstrates that the ML model becomes an integral part of the clinical workflow [12] and infrastructure, facilitating evidence-based decision-making and promoting data-driven solutions. This integration ensures efficient use of data and contributes to more effective and informed actions in managing health challenges.

**Data privacy and Security in ML solutions:** this indicator refers to how an ML solution safeguards sensitive and personal data from unauthorized access, hacking, and improper use. It assesses the robustness of the model's data protection measures and adherence to privacy regulations [13]. A high value for this indicator indicates that the ML solution prioritizes data privacy and security, adheres to regulations, and employs robust measures to protect sensitive information. Robust data privacy and security measures build trust among users and stakeholders, ensuring responsible and ethical use of data in ML applications in the health field.

**Model Reliability:** The availability of data greatly influences ML models' reliability. In the health field, datasets often contain only a small number of samples, highlighting the importance of having larger training datasets to ensure a reliable model. Machine learning models may struggle to generalize to new, unseen instances from a population different from the one used for training or an under-represented subpopulation [14]. This indicator aims to evaluate the reliability of the model by taking into account both the complexity of the model and the size of the dataset used for training. Linear models are generally considered more suitable for training with limited data, and shallow learning algorithms are preferred over deep ones. For instance, some studies have been conducted to establish the relationship between the number of features and the number of training samples [15].

However, employing deep architectures trained using a transfer learning approach that takes advantage of large open-source databases as source datasets is often possible. This allows for the transfer of knowledge to smaller, proprietary datasets.

**Reproducibility:** To be considered trustworthy, ML models must exhibit computational reproducibility [18]. To achieve reproducibility, researchers must offer comprehensive documentation of their experimental setup, which encompasses details about the dataset used, data preprocessing methods, model architecture, hyperparameters, and any specific algorithms or

frameworks utilized. Additionally, it is essential to share instructions on how to reproduce the results.

**Usability:** Machine learning models frequently provide only the probability of a prediction, posing a challenge for humans to incorporate the model into their decision-making process. Consequently, many machine learning algorithms lack usability, which refers to their ability to be efficiently utilized by humans to make informed decisions [17]. Regardless of the level of technical expertise needed for implementation, it is crucial for ML models to be accessible and capable of providing decision support, even to individuals without technical knowledge. Thoroughly documenting these models is essential to ensuring ease of use, which entails offering clear explanations of the system's inputs, instructions on preparing them, and guides on interpreting the output. This indicator aims to assess the user-friendliness of the developed model for end users.

## 5. Applicative Scenarios

In order to prove the usefulness of the proposed indicators, we report some applicative scenarios where the hitherto defined indicators could be used to formulate recommendations and give quantitative and qualitative responses to practical and current examples.

**Interoperable health and digital environmental data for primary prevention:** It aims to protect communities and the environment from high-intensity pollution generated by environmental emergencies or disasters, including the effects of climate change and natural hazards, by adopting an 'Assess, WArn & Response' (AWARE) approach in a coordinated inter-institutional effort in support of preparedness.

**A population-based digital approach to interoperate cancer registries, specialized clinical/pathology networks, and data flows:** It aims to develop digital functions in support of an advanced interoperable cancer surveillance system to investigate the risk factors associated with cancer occurrence, predict the risk of cancer development, and perform real-world analyses, while interoperating cancer registries with specialized pathology registries from clinical specialized networks, and data from the general population.

**Interoperating population-based registries and environment monitoring systems:** To develop an advanced interoperable cancer surveillance system powered by an innovative digital infrastructure



integrating data from cancer registries with data from environmental monitoring systems, allowing for the implementation of innovative community-based digital primary preventive strategies.

**Managing the effects of environmental exposures across the lifespan on health outcomes in different target populations:** It aims to protect community health across the lifespan in different target populations from long-term exposure to environmental pollution, following a coordinated inter-institutional effort and using suitable data mining solutions.

Monitoring lifestyles and health determinants in different settings and population targets through novel technological approaches for digital primary prevention: It aims to create a database in which the data collection system is based on digital technologies and to develop predictive models for the primary prevention of chronic non-communicable diseases.

**A multivariable model beyond the state of the art for predicting incident falls in community-dwelling older subjects:** A study could test whether older people's fall risk may be predicted more effectively than the current state of the art by adding information on spontaneous activity, sleep, and heart rate derived from wearable sensors for five days every three months. A new time-variant fall risk prediction model could be developed and tested in subjects recruited by hospitals and general practitioners.

**Implementing an interoperable web-based platform to support health surveillance against latent tuberculosis infection:** Tuberculosis (TB) prevention is a major goal in a teaching hospital setting. Because of the possible progression or reactivation of latent disease, the screening of both healthcare workers and students is an important issue in the TB control program.

**Prevention of falls and injuries:** It aims to define a set of measures and methods to identify subjects at high risk of falling. Muscle power (measured during isometric/isokinetic dynamometry and dynamic tests) and motor control deficits (estimated using digital twins and in silico methods) could be compared to and/or combined with muscle strength measures to better characterize/assess older individuals at risk of falling or to prevent future falls.

**Multi-drug Resistant Organisms (MDRO) Analysis and Surveillance System through automatized elaboration of laboratory data:** it aims to acquire, clean, and analyse healthcare data on multi-drug resistant organisms (MDRO) in healthcare facilities. The analysed data will be

showcased in an interactive dashboard, and performance benchmarks will be set for ongoing monitoring and improvement of MDRO health management and prevention.

**Digitalization of vaccination processes and integration with surveillance systems:** it aims to support with digital tools all the vaccination processes, from vaccination targeting to invitation and appointment reservation, and registration of vaccines' administration. Moreover, a digitalised function will be deployed to allow a post-vaccination follow-up. The vaccination databases and registries will be further interoperated with the institutional surveillance systems related to vaccine-preventable infectious diseases. Lastly, data mining techniques will be applied to monitor and analyse the impact of vaccination in accordance with data from the Surveillance System for respiratory diseases and other vaccine-preventable diseases (VPDs) and to evaluate vaccine effectiveness.

## 6. Case Study Example: cardiovascular disease risk prediction as a primary prevention strategy

CML-Cardio is a machine-learning model for primary prevention of cardiovascular disease [19]. Specifically, the model was trained to classify patients according to their risk of developing cardiovascular disease. In particular, a cascade model consisting of two stages was implemented to assess the risk of cardiovascular disease. The proportion of high-, intermediate-, and low-risk classes was highly unbalanced. Shallow learning algorithms, such as support vector machine, random forest, xgboost, etc., were used. Most of the previously proposed indicators can be applied and used.

**Impact on Health:** CVDs rank among the primary causes of mortality worldwide, accounting for 40% of global deaths [20]. Despite the potential for prevention and widespread awareness regarding the importance of adopting healthy habits, it is projected that by 2035, approximately 45% of individuals in the United States will experience a cardiovascular disease-related issue.

**Prospective Utility:** The Global Risk Score (GRS) enables the estimation of an individual's likelihood of developing CVDs within the next decade. By considering readily available factors such as smoking, diabetes, dyslipidemia, and hypertension, the GRS can be calculated to determine

community-based preventive measures for each individual. This approach allows tailored interventions based on an individual's risk profile.

**Economic Impact:** Even if several studies do not include an economic evaluation, implementing an application should not incur excessive costs, as it primarily relies on the knowledge of easily accessible parameters such as age, smoking status, blood pressure, and so on.

**Geographical Impact:** several works were carried out using a cohort of patients acquired in local countries, such as Brazil. To establish validity on a large scale, the cohort should include patients from different countries, even different continents.

**Existing Interventions:** At present, each factor (smoking, diabetes, etc.) plays a role in calculating the GRS for categorizing cardiovascular risk as low, intermediate, or high. However, the current manual calculation method is prone to errors by the specialist. To address this issue, implementing ML models provides valuable support by considering the interrelationships among variables, thereby enhancing the objectivity of the risk classification process.

**ML model evaluation:** The used model performance should be evaluated and assessed through cross-validation, accuracy, precision, recall, F1-score, and the area under the receiver operating characteristic curve (AUC-ROC). In addition, the procedure to evaluate the combined model (to assess the three-class classification) should be described. With the use of the confusion matrix, it is possible to verify the performance also in highly unbalanced scenarios.

**Metadata richness:** The Metadata processed by ML algorithms include features of age, blood pressure, diabetes, smoking information, cholesterol, body mass index (BMI), weight, height, and abdominal circumference. For example, in [19], the final dataset consisted of 71 subjects whose statistics on all features were collected in two different periods. Data acquisition was carried out in August 2016 and February 2018. This study allowed to verify changes in the measures that occurred in each subject over the years. The final dataset was also published on the Mendeley platform [21].

**Environmental-Health Correlation Estimation Accuracy:** it is impossible to directly extract the efficiency level of ML models in identifying and quantifying the Environmental-Health Correlation Estimation Accuracy based on disease risk level. In [19], only data pre-processing was addressed by advising the subjects with diabetes involved to change their habits to decrease the risk of CVD. It has been attributed as a limitation of this work the fact that the data used to train the ML models

come from a small population sample, being all employees of the same company and residing in the same geographical region, sharing the same climate, the same type of basic hygienic facilities, having similar eating habits.

**ML model robustness:** in [19], the results obtained through the nested cross-validation approach show that the MLP model performed best for all metrics considered, also presenting an AUC-ROC of  $0.91 \pm 0.03$  thus a greater stability of results, with a smaller standard deviation respect to the other ML models. However, the developed ML models should be further explored in other large sets of health data, in other populations, and for predictions of other disease outcomes.

**ML model scalability:** the work [19] did not provide information about the model's performance in relation to data growth or the scale of computational resources utilized. Therefore, evaluations on this indicator cannot be addressed in the context of the presented work.

**Health ML systems interoperability:** The prototype created was made available to the public after a period of use by the participating private company. However, the work did not address the interoperability of the ML prototype with pre-existing ML health systems and infrastructure related to cardiovascular disease. It did not provide information on how the ML model integrates with existing health systems, improves their capabilities, or facilitates seamless information exchange.

**Data privacy and Security in ML solutions:** To preserve privacy and protect the confidentiality of all subjects involved in the study, data were anonymized. Thus, all personal information in the final public dataset were removed, helping to ensure the confidentiality of the data used in the ML solution.

**Model Reliability:** The number of samples is minimal for training ML models. However, the number of features utilized is also limited, and the emphasis was placed on employing shallow learning algorithms rather than deep learning algorithms, which typically necessitate a more considerable amount of data. In addition, the metrics were evaluated by considering cross-validation, giving a more accurate estimate of generalization ability.

**Reproducibility:** data are public and easily accessible. The code is unavailable; however, the workflow should be easily reproducible (the Scikit-Learn Python library was used).

**Usability:** A particular focus on providing an interpretation of the findings allows for comparison with existing medical literature. Moreover, the development of a prototype application with an

intuitive interface, designed to be easily usable even by users with limited experience, was proposed.

## 7. References

- [1] <https://www.cdc.gov/std/Program/pupestd/Developing%20Evaluation%20Indicators.pdf>
- [2] <https://mypeer.org.au/monitoring-evaluation/indicators-for-evaluation/>
- [3] Wilkinson, Mark D., et al. "The FAIR Guiding Principles for scientific data management and stewardship." *Scientific data* 3.1 (2016): 1-9.
- [4] Kuhn, M., & Johnson, K. (2013). *Applied predictive modeling* (Vol. 26, p. 13). New York: Springer.
- [5] Hicks, S. A., Strümke, I., Thambawita, V., Hammou, M., Riegler, M. A., Halvorsen, P., & Parasa, S. (2022). On evaluation metrics for medical applications of artificial intelligence. *Scientific reports*, 12(1), 5979.
- [6] Müller, D., Soto-Rey, I., & Kramer, F. (2022). Towards a guideline for evaluation metrics in medical image segmentation. *BMC Research Notes*, 15(1), 1-8.
- [7] Tohka, J., & Van Gils, M. (2021). Evaluation of machine learning algorithms for health and wellness applications: A tutorial. *Computers in Biology and Medicine*, 132, 104324.
- [8] Liolios, K., Schriml, L., Hirschman, L. et al. The Metadata Coverage Index (MCI): A standardized metric for quantifying database metadata richness. *Stand in Genomic Sci* 6, 444–453 (2012). <https://doi.org/10.4056/sigs.2675953>
- [9] Su, PF., Sie, FC., Yang, CT. et al. Association of ambient air pollution with cardiovascular disease risks in people with type 2 diabetes: a Bayesian spatial survival analysis. *Environ Health* 19, 110 (2020). <https://doi.org/10.1186/s12940-020-00664-0>
- [10] Subbaswamy, Adarsh, Roy Adams, and Suchi Saria. "Evaluating model robustness and stability to dataset shift." International conference on artificial intelligence and statistics. PMLR, 2021.
- [11] Sierra-Sosa, Daniel, et al. "Scalable healthcare assessment for diabetic patients using deep learning on multiple GPUs." *IEEE transactions on industrial informatics* 15.10 (2019): 5682-5689.
- [12] Reddy, Sandeep, et al. "A governance model for the application of AI in health care." *Journal of the American Medical Informatics Association* 27.3 (2020): 491-497.

- [13] Papernot, Nicolas, et al. "Sok: Security and privacy in machine learning." 2018 IEEE European Symposium on Security and Privacy (EuroS&P). IEEE, 2018.
- [14] Nicora, G., Rios, M., Abu-Hanna, A., & Bellazzi, R. (2022). Evaluating pointwise reliability of machine learning prediction. *Journal of Biomedical Informatics*, 127, 103996.
- [15] Chalkidou, A., O'Doherty, M. J., & Marsden, P. K. (2015). False discovery rates in PET and CT studies with texture features: a systematic review. *PloS one*, 10(5), e0124165.
- [16] Weissler, E. H., Naumann, T., Andersson, T., Ranganath, R., Elemento, O., Luo, Y., ... & Ghassemi, M. (2021). The role of machine learning in clinical research: transforming the future of evidence generation. *Trials*, 22(1), 1-15.
- [17] Zytek, A., Liu, D., Vaithianathan, R., & Veeramachaneni, K. (2021). Sibyl: Understanding and addressing the usability challenges of machine learning in high-stakes decision making. *IEEE Transactions on Visualization and Computer Graphics*, 28(1), 1161-1171.
- [18] Heil, B. J., Hoffman, M. M., Markowetz, F., Lee, S. I., Greene, C. S., & Hicks, S. C. (2021). Reproducibility standards for machine learning in the life sciences. *Nature Methods*, 18(10), 1132-1135.
- [19] Oliveira, B. A. S., Castro, G. Z., Ferreira, G. L. M., & Guimarães, F. G. (2023). CML-Cardio: a cascade machine learning model to predict cardiovascular disease risk as a primary prevention strategy. *Medical & Biological Engineering & Computing*, 1-17.
- [20] Dagenais, G. R., Leong, D. P., Rangarajan, S., Lanas, F., Lopez-Jaramillo, P., Gupta, R., ... & Yusuf, S. (2020). Variations in common diseases, hospital admissions, and deaths in middle-aged adults in 21 countries from five continents (PURE): a prospective cohort study. *The Lancet*, 395(10226), 785-794.
- [21] Ferreira, Giovanna; Arruda, Caroline; Rodrigues, Clara; Isidoro, Gabriela (2023), "Cardiovascular Disease Risk Dataset", Mendeley Data, V1, doi: 10.17632/vhgyn5yk4g.1