



DARE

DIGITAL LIFELONG PREVENTION

CODE NO. PNC0000002

Spoke 2 Deliverable

S2.D4.3

Predictive algorithms

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



Deliverable information

Spoke number and title	Spoke 2 - Community-Based Digital Primary Prevention
WP number and title	WP 4 - Digital tools for Primary Prevention
Related task(s)	<p>Task 4.1- Predictive models for automatic disease surveillance system</p> <p>Task 4.2 - Use of Digital Technologies to Support Vaccination programs</p> <p>Task 4.3 - A sustainable and technological approach to large-scale prevention of falls and injuries</p> <p>Task 4.4 - Innovative digital tools for personalized cardiovascular primary prevention</p>
Lead beneficiary	FPG
Contributing beneficiaries	UNIBO, UCSC, UNIPA, IOR, IRCCS AOU BO, AUSL Romagna, ROMA1, UPCMI
Dissemination level	Public, fully open
Due date	15/06/2025
Actual date of delivery	14/06/2025
Author(s)	Roberta Pastorino (FPG), Alessandro Silvani (UNIBO), Claudio Costantino (UNIPA), Matteo Di Pumpo (FPG)
Contributors	Paolo Parente (ROMA1), Andrea Barbara (ROMA1), Luigi Russo (UCSC), Andrea Gentili (UCSC), XXXX
Quality Assurance	Walter Mazzucco (UNIPA), Alessio Signorello (Almaviva), Valeria Lukaj (Almaviva)

Document history

Version	Date	Author(s) /Reviewer(s) (Beneficiary)	Description
0.1	7/06/2025	Roberta Pastorino (FPG), Alessandro Silvani (UNIBO), Claudio Costantino (UNIPA), Matteo Di Pumpo (FPG)	First draft
0.2	10/06/2025	Alessio Signorello (Almaviva), Valeria Lukaj (Almaviva)	Revision
0.3	12/06/2025	Walter Mazzucco (UNIPA)	Final Revision
0.4	14/06/2025	Roberta Pastorino (FPG)	Final Revision

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

Publishable summary	6
1. Introduction.....	6
2. Predictive models for automatic disease surveillance system.....	9
2.1. Predictive models for automatic disease surveillance system from the development of a Data Lakehouse platform to clinical pathway classification, monitoring and forecasting disease evolution and the impact of climate changes on the hospitalization - IRCCS AOUBO and UNIBO	9
2.2. Hand Hygiene: Practices, Techniques, and Knowledge Among Healthcare Professionals. Field Evaluation of Best Practices. “HyPTeK”- FPG.....	14
3. Use of Digital Technologies to Support Vaccination programs	17
3.1. Caring for frail patients through vaccination – CAREVAX -FPG	17
3.2. Empowerment for vaccinating Communities: Small world networks approach – ROMA 1.....	23
3.3. Digi-Vax: digitalization of vaccination processes and integration with surveillance systems - UNIPA	26
4. A sustainable and technological approach to large-scale prevention of falls and injuries	29
4.1. BONESTRENGTH: An <i>in-silico</i> clinical trial (ISCT) technology to assess the efficacy of intervention strategies for the prevention of hip fractures - UNIBO	29
4.2. Muscle power and motor control degradation are better predictor of falls than muscle strength in the aging population - IOR and UNIBO	33
4.3. DARE-FALLSPREDICT: development of a multi-variable model beyond the state of the art for estimating the risk of falling in older people - UNIBO and AUSL ROMAGNA..	36
5. Innovative digital tools for personalized cardiovascular primary prevention - UCSC ...	44
5.1. Personalised HeartCare (PHC): innovative approaches for personalized primary prevention of cardiovascular diseases (CVDs)	47

5.2. Evaluation of Polygenic Risk Score for epithelial OVarian cancEr risk prediction and clinical outcomes in an Italian population: the PROVE study.....51

5.3. Integrated Genetic Risk Models (MIG) with Digital Solutions to Transform Breast Cancer Prevention: Assessment of Health and Care Impact 55

5.4. PRE-PDAC, Evaluation of Polygenic Risk scorE for Pancreatic Ductal AdenoCarcinoma risk prediction: a case-control study 58

Publishable summary

The DARE project is an innovative initiative aimed at enhancing disease prevention through the use of digital technologies and data-driven approaches. **Deliverable 4.3** presents the pilot activities developed within Work Package 4 (WP4), "Digital Tools for Primary Prevention," **with a specific focus on the predictive algorithms adopted to support personalized and proactive healthcare interventions.** Each pilot addresses a distinct public health challenge and leverages predictive modeling to identify at-risk populations, anticipate health events, and guide timely and targeted preventive measures:

Task 4.1 implements predictive models for automatic disease surveillance, using big data analytics to detect trends in hospitalizations and environmental exposures linked to climate change.

Task 4.2 supports vaccination strategies by applying data mining techniques to immunization records and developing predictive algorithms to identify individuals most likely to miss scheduled vaccinations.

Task 4.3 applies predictive algorithms to data from wearable sensors to improve the prediction of fall risk in older adults, enhancing current assessment tools.

Task 4.4 employs polygenic risk scores and behavioral indicators to guide personalized prevention strategies: lifestyle-based interventions for cardiovascular disease, and genetics-informed preventive approaches for cancer, which may include clinical surveillance or risk-reducing procedures such as prophylactic surgery.

The deliverable details the methodological approaches used to build and validate these predictive algorithms, the types of data sources integrated, and the expected impact on preventive healthcare. The insights gained from these pilot projects contribute to a broader understanding of how predictive analytics can be embedded into routine care pathways to support more efficient, equitable, and personalized prevention strategies.

1. Introduction

The DARE project represents a groundbreaking initiative in research, emphasizing innovation and technological progress, particularly within the healthcare system. Spoke2, a pivotal component of the project, is dedicated to driving transformation and innovation

in primary prevention by leveraging new technologies or reimagining existing ones with enhanced functionalities.

Within Spoke2, Work Package 4 (WP4), titled "Digital Tools for Primary Prevention," is focused on designing and implementing pilot projects that promote primary prevention activities through technological advancements underpinned by scientific evidence. The overarching aim is to create scalable and adaptable strategies suitable for diverse contexts. WP4's objectives include fostering collaboration between primary healthcare providers and hospitals through experimental protocols and innovative digital tools for data collection, management, and analysis. The activities encompass key areas such as disease surveillance, vaccination programs, cardiovascular risk profiling, and fall prevention strategies.

Overview of WP4 Tasks:

Task 4.1 - Predictive Models for Automatic Disease Surveillance Systems

This task aims to develop advanced monitoring methods and predictive models, integrating them through robust data analysis and extensive big data collection. A comprehensive set of statistical tools will be created to identify key turning points in daily hospitalization trends across different spatial and temporal scales. These tools will also facilitate the monitoring of environmental exposures linked to climate change, providing actionable insights for public health planning. Another aim is to assess hand hygiene practices, techniques and knowledge among healthcare workers using digital hand hygiene scanners for objective assessment. The project aims to measure baseline compliance, provide targeted training and evaluate short and long-term improvements. By integrating behavioural assessment with real-time feedback, the study supports continuous professional development and infection prevention. This aims to improve adherence to WHO guidelines and contribute to antimicrobial stewardship efforts.

Task 4.2 - Using Digital Technologies to Support Vaccination Programs

Task 4.2 focuses on devising innovative digital strategies to enhance coordination between community and hospital sectors in executing vaccination campaigns for specific vaccine-preventable diseases. Emphasis is placed on targeting the most vulnerable at-risk populations and integrating vaccination initiatives with existing surveillance systems for vaccine-preventable diseases (VPDs). The task explores pioneering methodologies such as:

- Data mining across Regional/National vaccination registries, hospital or ER discharge records, and VPD surveillance systems.

- Interoperability of Regional vaccination registries with Personal Health Records and with existing app that contain this data (e.g., "IO app").
- Implementing computerized reservation and reporting systems tailored for at-risk groups.
- Disseminating accurate vaccine safety and efficacy information through social media platforms.
- Employing innovative tools to address and mitigate vaccination hesitancy determinants.

Task 4.3 - Sustainable and Technological Approaches to Large-Scale Fall and Injury Prevention

This task addresses the limitations of fall risk assessment and prevention due to small sample sizes and study variability. By combining in-silico trial technologies with specialized prospective studies, it utilizes cost-effective wearable sensor technologies for widespread prevention efforts. Different predictive algorithms are applied to formulate robust risk models, enabling a more comprehensive and effective approach to preventing falls and related injuries.

Task 4.4 - Innovative Digital Tools for Personalized Cardiovascular Primary Prevention

Task 4.4 focuses on personalized prevention strategies informed by polygenic risk scores, either independently or in combination with digital technologies. In the cardiovascular domain, a community trial will assess the impact of innovative primary prevention interventions targeting lifestyle modification. This involves tailored motivational strategies to promote and sustain healthy behaviors at the population level. In the oncology context, genetic risk stratification is applied to cancers such as breast, ovarian, and pancreatic cancer, to inform individualized preventive approaches. These may include enhanced clinical surveillance or risk-reducing procedures, such as prophylactic surgery, for individuals at high genetic risk.

Deliverable 4.3 provides a detailed description of each pilot project, highlighting the predictive modeling approaches adopted, each tailored to the specific context, data sources, and application domain. Particular attention is given to the relevance of the variables selected for each model, as well as to the methods used for training and validating the algorithms. The following sections present each task along with its modeling strategy and methodological framework.

2. Predictive models for automatic disease surveillance system

2.1. Predictive models for automatic disease surveillance system from the development of a Data Lakehouse platform to clinical pathway classification, monitoring and forecasting disease evolution and the impact of climate changes on the hospitalization - IRCCS AOUBO and UNIBO

Introduction

The premise for the construction of predictive models is the design and implementation of an integrated system for the storage, management, analysis, and availability of the information assets of a healthcare research institute, to support administrative and research activities.

The system to be implemented will consist of:

1. A Data Lakehouse, a system that combines the features of a Data Lake and a Data Warehouse, allowing the storage and analysis of large volumes of both structured and unstructured data within a single platform, thereby facilitating more flexible and efficient access to information.
2. An advanced development environment for descriptive, diagnostic, predictive, and prescriptive analyses, integrated with a set of APIs for data access via DataFrames and for interoperability with modern data science platforms. This will enable the development of predictive models, numerical simulations, and combinatorial optimization.

The goals the project aims to achieve through this supply are:

- To integrate data from all applications used at IRCCS AOU Bologna, including both core departmental systems of a Hospital Information System (HIS) and applications related to research activities. This also includes the ability to store and manage unstructured data from external applications that are relevant to healthcare processes of interest.
- To enable the creation and management of both static and dynamic dashboards for retrospective and real-time analysis, directly connected to the organization's data lakehouse.
- To allow access to and processing of data using open-source advanced analytics systems, for integration with Artificial Intelligence, Predictive and Prescriptive Analytics models and algorithms developed within IRCCS AOU Bologna.

Data Collection for Algorithm Development

The system will be implemented experimentally at the IRCCS AOU Bologna and will follow the conceptual framework shown in the figure below.

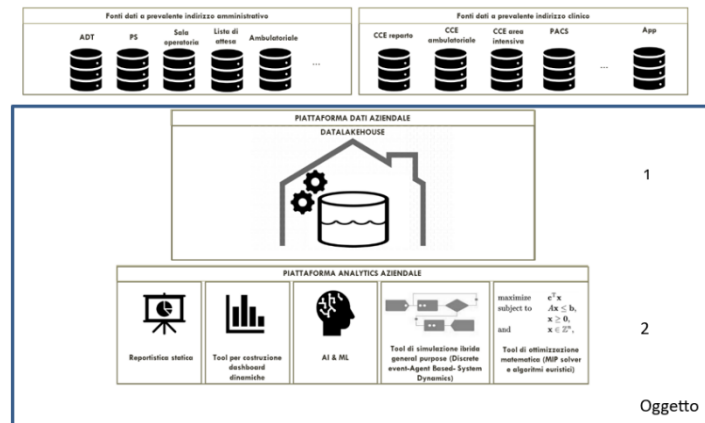


Figura 1

Figure 1. Conceptual framework of the system.

Technical details of the Algorithm and its Applications

From the modeling perspective, a tripartite predictive framework for hospitalization time series and health risk will be developed, composed of the following modules:

- Epidemiological dynamical model;
- Hazard-based model for acute events and hospital admissions;
- Joint monitoring and forecasting methods for automatic disease surveillance.

A compartmental model will be developed based on delay differential equations, to forecast the evolution of epidemic outbreaks at metropolitan level. The model aims to cope with the problem of predicting and controlling the infected individuals from the risk compartments since the number of hospitalized individuals is derived using a hospitalization probability that depends on the original risk class of the population. The susceptible individuals will be divided into different compartments correlated with the hospitalization probability (i.e. the probability to develop a serious disease) that is inferred from the correlation among the admissions to the emergency room and the hospitalization in different departments. This model will be informed by data extracted from the data Lakehouse developed by IRCCS AOU Bologna. The compartments may depend on social features (like age, job type) or on the comorbidity of the disease with other pathologies that increases the probability of developing a serious disease. The effect of parameter fluctuations will be estimated, and the social activity parameter will be quantified computing the susceptibility of the system in presence of small random fluctuations and proxy available data sets (e.g. the traffic open data in Bologna) that correlate with the social activities in the city will be used. The epidemic spreading and the impact on the

hospitalizations will be studied by performing simulations according to the following scheme (see Fig. 2):

1. after a contact with an unreported infected individual, a susceptible individual becomes an unreported infected individual with a contagion probability;
2. an unreported infected individual may become an infected hospitalized individual or recovered individual after a time defined by a random variable;
3. a hospitalized individual is removed from the active population and will enter the recovered compartment after the healing time;
4. a recovered individual has a probability to return susceptible after a certain time scale;
5. if a vaccine is available, a susceptible is transferred to the vaccinated compartment and returned to the susceptible compartment with a given probability after a certain time.

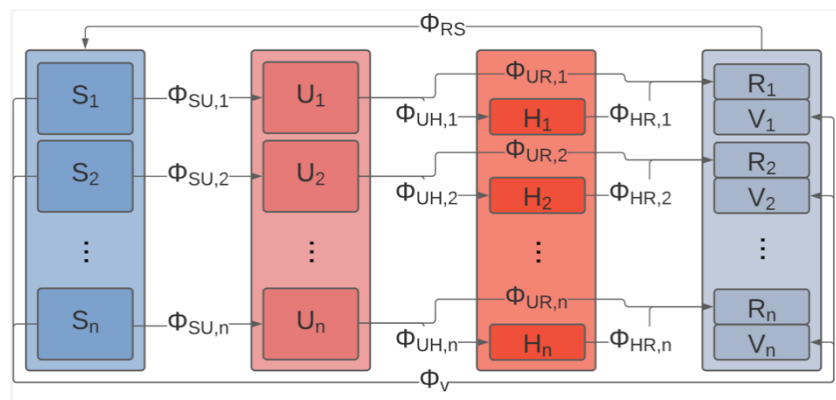


Figure 2. Scheme of the compartmental model for epidemic spread simulations, where the indices 1,...,n indicate the multiple social compartments and the arrows the corresponding flows.

In the field of survival analysis, the study of the expected amount of time before a given event occurs, a common class of models relies on the specification of a hazard function, dictating the event rate after a given time, provided that the event has never happened until then. Complementing the other forecasting approaches, a model will be proposed based on the concept of hazard function, outlined in the following. The lifetime variable of the survival process will be taken as the time since the beginning of the heat wave, marked by a raised temperature. The hazard function will then model the rate at which, after a certain time of exposure to the heat wave, patients will start to show symptoms, be admitted to the ER and in a fraction of the cases hospitalized. This model will also be stratified in several risk categories, characterized by different hazard functions, accounting for the fact that fragile individuals such as infants or the elderly population are generally affected earlier and more seriously than the other ones. A refinement of the model would imply introducing a continuous dependence on the temperature of the rate at which

individuals enter the survival process, and a further implementation should consider also the impact of hospitalizations and ER care of heat wave related affections on the hospital system.

A set of statistical and computational tools are proposed for disease surveillance, combining predictive models and monitoring methods. The main objective is to support the analysis of daily hospitalizations and other quantities related to health data, including admissions to the emergency unit and other specific departments, number of detected infections, and deaths. The proposed methodologies will allow researchers to evaluate the evolution of a phenomenon of interest, produce short-term forecasts, detect eventual changes in trend, and examine the impact of exogenous variables such as meteorological events and the presence of air pollutants. The methods will be integrated into the data lakehouse infrastructure provided by IRCCS AOU Bologna, enabling real-time analysis of health data to support the prompt implementation of policies in response to disease evolution and predicted trends. The final deliverable will be an integrated model on a platform that provides visualization tools for the simulations, with the capability to automatically update information daily.

The necessary tools are constructed combining advanced models for time series for count data and process control charts, that have proven to be valuable methods to provide an effective description of health-related phenomena (COVID-19, influenza-like diseases, heat-related illnesses, etc.). First, short-term predictions over a 7-day horizon are computed through an ensemble modelling approach, which exploits and combines forecasts from multiple models. Several works on medical data and other fields showed that ensemble approaches often give more accurate forecasts than single models, as they integrate the strengths of different models, while compensating for the individual weaknesses and producing more robust predictions. The proposed ensemble relies on multiple models, including flexible regressions with cubic splines, generalized additive models, Bayesian structural time series, and integer-valued generalized autoregressive conditional heteroscedasticity models. Ensemble predictions are computed as weighted averages of the individual models' predictions, with weights determined according to performance on the last available counts. Some weights are set to zero, so that only the models with the best performance are included in the ensemble. An immediate extension is the inclusion into the ensemble of predictors of interest. The resulting procedure is extremely general and flexible, allowing for any suitable choice of the count time series under analysis, the prediction window, and the covariates.

Subsequently, results are employed to monitor the phenomenon and identify potential change points using a two-step process control procedure that relies on nonparametric Sheward-type quality control charts. The counts will be modeled against time (calibration

step), then the subsequent counts will be estimated and the residuals monitored through a Cramer-von Mises Change Point control chart (monitoring step). In this setting, using the ensemble predictions within the procedure often allows for an earlier detection of change points.

The performance of this methodology has been evaluated, among others, in the analysis of daily counts of COVID-19 hospitalizations in Italian regions. Results confirm that the proposed methods can provide reliable short-term predictions and discern substantive causes of directional variation from random fluctuations around a baseline trend. Indeed, the performance of individual models varies significantly between the periods, while the construction of the ensemble mitigates extreme behaviors and tends to produce more stable results. Finally, the two-step process control procedure effectively detects change points and, in many cases, the integration of ensemble predictions leads to an earlier detection. The methodology and its applications appear in [1]. They were presented at the 52nd Scientific Meeting of the Italian Statistical Society (June 2024, Bari, Italy), the 32nd International Biometric Conference (December 2024, Atlanta, Georgia, United States), and AI for health and well-being @UNIBO: innovation in PNRR and PNC projects (February 2025, Bologna, Italy).

Future analyses will consider the data provided by the AOC IRCCS Bologna, with a focus on admissions to different units of the hospital and the health impact of environmental exposures, such as meteorological events and levels of air pollution. Furthermore, additional data will be incorporated such as individual characteristics of patients (socio-demographic, clinical, geographical, etc.) to explore the impact of diseases on different compartments of the population.

References

- [1] Vesely, A.; Roli, G.; Scagliarini, M.; Miglio, R., [*An Ensemble Method for Disease Surveillance*](#), in: Methodological and Applied Statistics and Demography IV, Springer Cham, «ITALIAN STATISTICAL SOCIETY SERIES ON ADVANCES IN STATISTICS», 2025, pp. 649 - 654.

2.2. Hand Hygiene: Practices, Techniques, and Knowledge Among Healthcare Professionals. Field Evaluation of Best Practices. “HyPTeK”- FPG

Introduction

Healthcare-Associated Infections (HAIs) continue to pose a major threat to patient safety, contributing significantly to hospital-related morbidity, mortality, and antibiotic overuse. The World Health Organization (WHO) has long identified correct hand hygiene as the most effective measure to prevent HAIs [1-8]. Despite international guidelines and institutional protocols, adherence to proper hand hygiene techniques among healthcare professionals remains suboptimal, often due to inadequate training, lack of feedback, and subjective assessment methods.

The HyPTeK project responds to this challenge by integrating an innovative digital assessment tool—the Semmelweis Scanner—into routine clinical practice at the Fondazione Policlinico Universitario A. Gemelli IRCCS (FPG). The aim is to objectively measure the correctness of hand hygiene techniques and assess the impact of targeted educational interventions over time.

This initiative is particularly timely in the broader context of antimicrobial stewardship, where reducing the incidence of HAIs plays a direct role in limiting unnecessary antibiotic prescriptions and, consequently, the emergence of antimicrobial resistance (AMR). By promoting evidence-based training, continuous monitoring, and real-time feedback, HyPTeK contributes to a culture of prevention and quality improvement in infection control.

Data Collection for Behavioral Assessment

The HyPTeK study is structured as a single-arm, prospective interventional trial and will enroll 118 healthcare professionals from various Operational Units (OUs) at FPG. The study population is drawn from a broader institutional cohort of over 6,700 healthcare workers and residents from Università Cattolica del Sacro Cuore (UCSC). Participant data will be collected and processed using in-house software in compliance with institutional data governance standards.

The evaluation follows a four-step timeline:

- **T0 (Baseline – June 2025):** Participants are tested using the Semmelweis Scanner. Hand hygiene performance is objectively scored, with 100% surface coverage classified as “appropriate”.
- **T1 (One month after T0):** Participants are invited to attend an educational seminar on correct hand hygiene practices, based on WHO guidelines. Non-attendees are excluded from follow-up.
- **T2 (One month after T1):** A second scanner-based assessment is performed to evaluate short-term improvements following the intervention.
- **T3 (Three months after T2):** A final assessment is conducted to determine long-term retention of hand hygiene practices.

Collected data will be stratified by professional category and clinical unit, enabling comparative analysis of hand hygiene adherence across different healthcare roles.

Sample size calculation was based on an assumed baseline adherence of 80%, with the goal of detecting an increase to 90% post-intervention. Setting $\alpha = 0.05$ and $\beta = 0.20$, and assuming a 10% dropout rate, the minimum required sample is 118 participants.

Technical Details of the Assessment Tool and Data Platform

At the core of the HyPTeK project lies the Semmelweis Scanner, a digital instrument designed to provide real-time, objective feedback on hand hygiene technique. Healthcare workers are instructed to apply a fluorescent gel and then place their hands under the scanner, which uses UV light and image recognition to detect covered and uncovered skin surfaces. Results are immediately displayed on screen, with percentages of correctly and incorrectly covered areas, and visual highlights of missed zones in red.

The scanner is able to identify individual users according to professional role and operating unit, by means of special recognition cards, which can be customised by type of operator or even by individual operator, allowing precise monitoring of compliance patterns. This data-driven approach eliminates the variability of human observation and introduces a standardized metric for quality control in hand hygiene.

In-house software infrastructure supports both local data storage and processing. Collected results are anonymized and stored securely in compliance with GDPR regulations. The system allows continuous tracking of performance over time, enabling the evaluation of educational impact and long-term behavioral change.

Ultimately, the HyPTeK model is designed to be scalable and replicable, offering a modular framework that can be integrated into institutional infection prevention strategies. By combining behavioral science, digital diagnostics, and clinical education, the project aspires to elevate hand hygiene from routine task to strategic intervention in antimicrobial resistance prevention.

References

- [1] P. Bolton and T. J. McCulloch, “The evidence supporting WHO recommendations on the promotion of hand hygiene: A critique,” *BMC Res Notes*, vol. 11, no. 1, Dec. 2018, doi: 10.1186/S13104-018-4012-3.
- [2] “Making health care safer II: an updated critical analysis of the evidence for patient safety practices - PubMed.” Available: <https://pubmed.ncbi.nlm.nih.gov/24423049/>
- [3] “Global report on infection prevention and control,” 2022.
- [4] “Infezioni correlate all’assistenza - EpiCentro - Istituto Superiore di Sanità.” Accessed: Jun. 05, 2025. [Online]. Available: <https://www.epicentro.iss.it/infezioni-correlate/>
- [5] “Sorveglianza Europea Mediante Prevalenza Puntuale Delle Infezioni Correlate All’assistenza E Sull’uso Di Antibiotici Negli Ospedali Per Acuti”.
- [6] C. Suetens et al., “Prevalence of healthcare-associated infections, estimated incidence and composite antimicrobial resistance index in acute care hospitals and long-term care facilities: Results from two european point prevalence surveys, 2016 to 2017,” *Eurosurveillance*, vol. 23, no. 46, Nov. 2018, doi: 10.2807/1560-7917.ES.2018.23.46.1800516,.
- [7] B. Allegranzi and D. Pittet, “Role of hand hygiene in healthcare-associated infection prevention,” *Journal of Hospital Infection*, vol. 73, no. 4, pp. 305–315, Dec. 2009, doi: 10.1016/J.JHIN.2009.04.019.
- [8] “WORLD ALLIANCE FOR PATIENT SAFETY LINEE GUIDA OMS SULL’IGIENE DELLE MANI NELL’ASSISTENZA SANITARIA (BOZZA AVANZATA) SFIDA GLOBALE PER LA SICUREZZA DEL PAZIENTE 2005-2006 Cure Pulite Sono Cure Più Sicure,” 2006, Accessed: Jun. 05, 2025. [Online]. Available: <http://www.agreecollaboration.org/pdf/agreeinstrumentfinal.pdf>

- [9] D. Pittet, B. Allegranzi, and J. Boyce, “The World Health Organization Guidelines on Hand Hygiene in Health Care and Their Consensus Recommendations,” *Infect* 10.1016/S0212-6567(14)70080-0.
- [10] V. Mouajou, K. Adams, G. DeLisle, and C. Quach, “Hand hygiene compliance in the prevention of hospital-acquired infections: a systematic review,” *Journal of Hospital Infection*, vol. 119, pp. 33–48, Jan. 2022, doi: 10.1016/J.JHIN.2021.09.016,.
- [11] “(PDF) Knowledge, habits and attitudes of health care workers about hand hygiene.” Accessed: Jun. 05, 2025. [Online]. Available: https://www.researchgate.net/publication/272179039_Knowledge_habits_and_attitudes_of_health_care_workers_about_hand_hygiene
- [12] D. Silva, O. Andrade, and E. Silva, “Perspective of health professionals on hand hygiene,” *Aten Primaria*, vol. 46, no. S5, pp. 135–139, Nov. 2014,
- [13] E. J. Septimus, “Antimicrobial Resistance: An Antimicrobial/Diagnostic Stewardship and Infection Prevention Approach,” *Medical Clinics of North America*, vol. 102, no. 5, pp. 819–829, Sep. 2018, doi: 10.1016/j.mcna.2018.04.005.
- [14] D. Kubde, A. K. Badge, S. Ugemuge, and S. Shahu, “Importance of Hospital Infection Control,” *Cureus*, vol. 15, no. 12, Dec. 2023, doi: 10.7759/CUREUS.50931.Control Hosp Epidemiol, vol. 30, no. 7, pp. 611–622, Jul. 2009, doi: 10.1086/600379,.

3. Use of Digital Technologies to Support Vaccination programs

3.1. Caring for frail patients through vaccination – CAREVAX -FPG

Introduction

Vaccine-preventable diseases (VPDs) remain a leading cause of morbidity, mortality, and healthcare costs, particularly among frail populations with chronic diseases, disabilities, or advanced age. The 2022–2025 Italian National Vaccine Prevention Plan (PNPV) [1] recognizes this vulnerability and calls for more proactive, digitally enabled strategies to enhance vaccine uptake. Despite some advances—such as the rollout of regional vaccination registries and hospital-based immunization initiatives—systemic challenges persist. These include fragmented hospital-community integration, disparities in digital infrastructure across regions, and the absence of automated tools to identify at-risk patients in real time.

The COVID-19 pandemic served as a catalyst for innovation, accelerating the adoption of in-hospital vaccination models and digital booking systems. However, vaccination coverage gaps remain, and digital registries often lack completeness or interoperability. Mission 6 of Italy's National Recovery and Resilience Plan (PNRR) seeks to address this by prioritizing digital health transformation, including the ongoing development of a National Vaccine Registry.

CareVax responds to these challenges by implementing a digitally integrated, hospital-based vaccination model that targets frail patients using clinical and immunization data [2]. The initiative aims to optimize access and trust in vaccination pathways by aligning eligibility screening with national guidelines and embedding it directly within hospital infrastructure. A pilot study is currently underway at Fondazione Policlinico Universitario A. Gemelli IRCCS (FPG), involving five departments (dermatology, geriatrics, gastroenterology, nephrology, and oncological gynecology), with nearly 100 patients recruited to date.

The project's overarching goal is to evaluate the feasibility, effectiveness, and scalability of a semi-automated alert system capable of identifying vaccine-eligible individuals and supporting personalized immunization strategies within a hospital-territory care continuum.

Data Collection for Algorithm Development

The algorithm underpinning CareVax relies on the integration of multiple data streams, primarily sourced from TrackCare, the institutional electronic health record system at FPG. TrackCare provides demographic data, hospitalization details, diagnostic codes, exemption codes, and relevant clinical histories. These are essential for establishing a patient's risk profile and eligibility for specific vaccines.

These hospital records are then cross-referenced with the Lazio Regional Vaccine Registry to assess vaccination history and identify missed opportunities. However, significant challenges persist: the national vaccine registry is still under development, and regional systems are not fully interoperable. Consequently, the system currently excludes patients with medical residence outside the Lazio region, and many older patients have incomplete

historical records due to a lack of digital documentation. Manual chart reviews and anamnestic checks by clinicians remain necessary to address these gaps temporarily. Data collection and management occur via REDCap, a GDPR-compliant platform hosted at FPG, ensuring data security, traceability, and interoperability. A dedicated pilot phase involving 10 patients was conducted to test the technical feasibility of the pathway, confirm data availability and flow, and validate the functioning of the eCRF and alert systems. Data are anonymized and stored according to best practices in data governance, with access restricted to credentialed users. Ongoing quality checks use the ACCIT framework to monitor completeness, accuracy, and data integrity across all collection points.

Technical details of the Algorithm and its Applications

The core of CareVax is a rule-based decision tree algorithm, developed in SAS and fully aligned with the eligibility criteria defined in the 2022–2025 Italian National Vaccine Prevention Plan. The rules encoded within the algorithm are directly derived from the recommendations and risk stratification criteria outlined in the PNPV, ensuring consistency with national public health priorities and clinical guidelines. The rule-based decision tree algorithm is developed in SAS and validated using a blinded review of mock electronic health records. The validation process demonstrated 100% agreement between the algorithm's vaccine eligibility assessments and those made by a panel of physicians, confirming its clinical accuracy. This validation step represents a key milestone in ensuring the algorithm's safe application in real-world settings.

The algorithm operates on structured inputs drawn from TrackCare, including:

- Age and time of year (e.g., influenza season),
- ICD9-CM codes indicating chronic diseases,
- Exemption codes for long-term health conditions,
- Indicators of recent clinical activity (e.g., hospital visits or therapies),
- Surrogate indicators (e.g., department admissions for pregnancy detection).

When a patient meets eligibility criteria for any of the target vaccines—SARS-CoV-2, Herpes Zoster, Influenza, *Streptococcus pneumoniae*, or Hepatitis B—the system populates the REDCap interface with the relevant vaccination recommendation. Physicians are then prompted to review the patient's profile, validate the algorithm's suggestions, and initiate contact for vaccine offering.

Although full automation is the long-term goal, current workflows include manual checks against the Lazio Regional Registry due to existing limitations in interoperability. The system is designed to adapt to future integration with the national registry, at which point automation will be extended to non-resident patients and eliminate the need for intermediary validation steps.

Ultimately, the algorithm is envisioned as a modular, embedded tool within hospital management systems. It supports clinicians with real-time decision support, facilitates direct scheduling of vaccinations, and contributes to broader objectives in public health automation and digital health equity. As digitization across public administration progresses, CareVax will require fewer manual interventions, making it more scalable and impactful across various care settings. The following figure 3 illustrates the complete patient access flowchart, detailing the utilization of the algorithm and associated data processes.

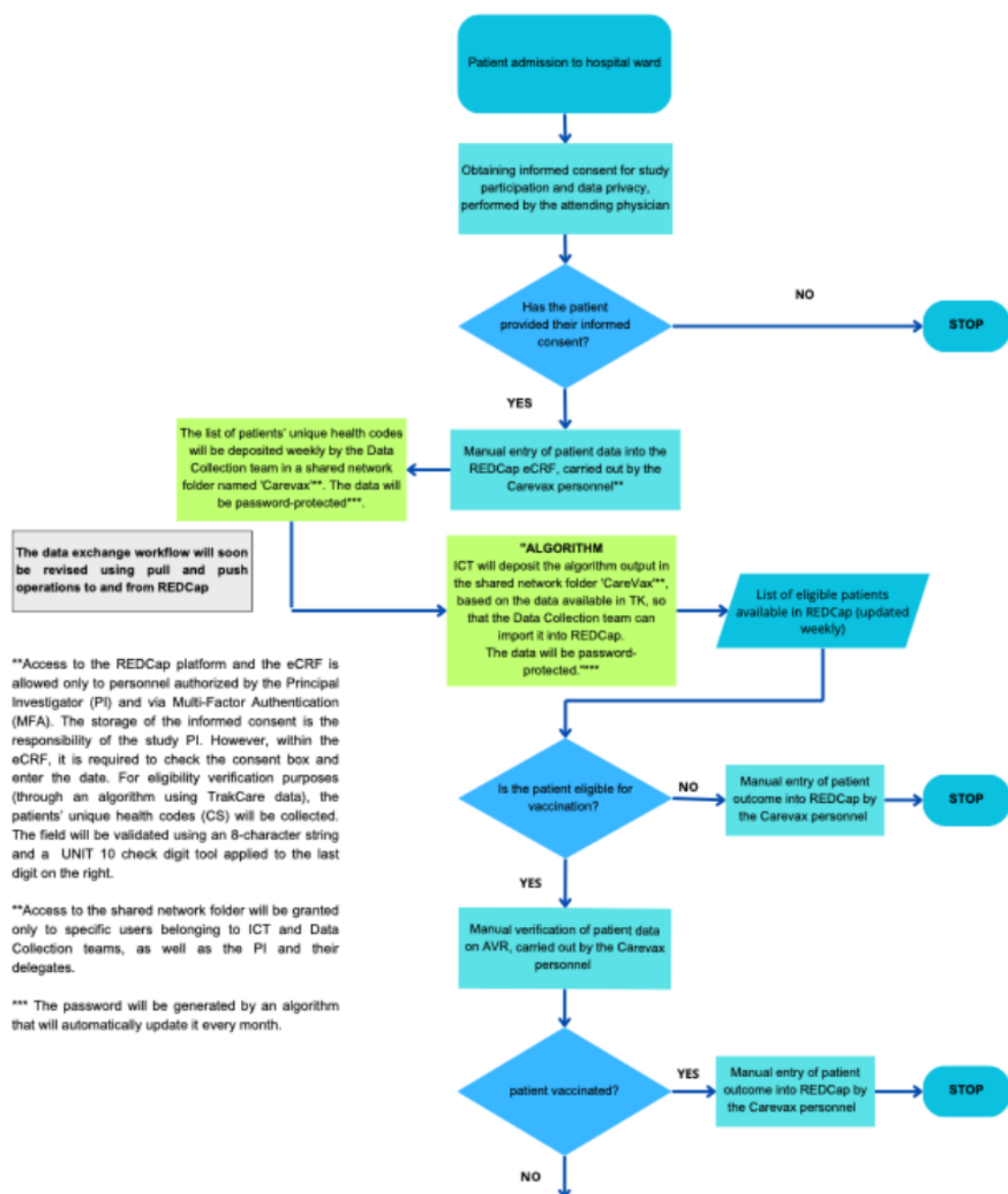






Figure 3. Flowchart of Patient Access and Management in the CAREX Study.

References

- [1] Gazzetta Ufficiale. (2024): Ministero della Salute Piano Nazionale Prevenzione Vaccinale 2023-2025 Available at: (<https://www.gazzettaufficiale.it/eli/id/2023/08/21/23A04685/sg>).
- [2] Lontano A, Regazzi L, Tona DM, Di Pumpo M, Porcelli M, Cacciuttolo MG, Parente P, Gasbarrini A, Grandaliano G, Panocchia N, Lopetuso L, Pasciuto T, Cadeddu C, Bruno S, Laurenti P, Pascucci D, Pastorino R. Digital integration between hospitals and local health authorities for enhanced vaccination coverage among frail patients: the CareVax study protocol. *Front Public Health*. 2025 Jan 30;13:1490244. doi: 10.3389/fpubh.2025.1490244. PMID: 39949559; PMCID: PMC11822475.

3.2. Empowerment for vaccinating Communities: Small world networks approach – ROMA 1

Introduction

The EVACS pilot project (Empowerment for Vaccination in Communities According to the Small World Networks Model) addresses the challenge of identifying and engaging hard-to-reach groups (HRGs) with the aim of increasing vaccination coverage, promoting

prevention, reducing health inequalities, and fostering community-centered healthcare. The initiative responds to the priorities outlined in the Italian National Vaccine Prevention Plan (PNPV) 2023–2025 and is aligned with the National Recovery and Resilience Plan (PNRR), which promotes the adoption of innovative strategies to address social vulnerability and health poverty.

Small world networks—such as informal communities in occupied buildings, nomadic camps, or marginalized school populations—tend to have strong internal ties but weak connections with institutional health services, making traditional vaccination campaigns less effective. EVACS proposes an integrated, digital, and community-based model that leverages social media sentiment analysis and local community structures to enhance equity and engagement in vaccination efforts.

Data Collection and Algorithm Development

The EVACS project foresees the integration of regional and national health platforms with tools capable of analysing digital content related to public perceptions of vaccines, with the goal of identifying signs of exclusion or vaccine hesitancy. Public content on social media will be analysed through Natural Language Processing (NLP) algorithms to detect negative sentiment and clusters of resistance to vaccination.

In addition, the system will monitor the response of small world populations to prevention campaigns and the implementation of the new European Harmonized Number (116117) and its related functionalities.

Alongside this digital component, territorial data sources will also be utilized—such as registries from Community Health Centers (Case della Comunità), school-based prevention programs, and third-sector services—in order to strengthen data triangulation and support the planning of targeted intervention strategies. These data will be harmonized with data collected by the Regional Prevention Department to cluster the population and identify not only the target groups but also their expressed needs.

The algorithmic system will adopt an ambispective observational design, combining retrospective data (e.g., vaccination registries and digital trends) with prospective monitoring. The ultimate goal is to enable timely identification and personalized outreach to the most vulnerable communities.

Technical Details of the Model and its Application

At the core of the project is the development of an electronic platform for mapping vaccination equity and enabling community empowerment, structured into several components:

- **Sentiment analysis module:** will collect and analyse public reactions across major social media platforms (including Facebook, Instagram, X/Twitter, TikTok, YouTube), monitoring comments, posts, hashtags, and viral content related to vaccination. The analysis, based on artificial intelligence and Natural Language Processing (NLP), will identify patterns of hesitancy, misinformation, or distrust, and will map them by territory, social group, and prevalent themes.
- **Targeting algorithm:** will integrate data extracted from social media with those from vaccination registries and local services in order to identify communities characterized by low vaccination coverage or significant cultural, linguistic, or digital barriers.
- **Personalized response system:** based on the sentiment analysis findings, dedicated communication campaigns will be designed and tested. These campaigns will be co-created with local stakeholders (healthcare workers, cultural mediators, schools, third-sector organizations) and disseminated through both digital and community channels. The campaigns will aim to overcome the identified barriers (e.g., misinformation, distrust, access difficulties) using targeted messages, visual storytelling, peer-to-peer testimonials, and multilingual materials.
- **Personalized linkage-to-care:** systems for vaccination notifications and patient navigation will be activated, tailored to the characteristics of the target groups and based on proximity models (e.g., mediators in neighborhoods, information desks in Community Health Centers, simplified booking via SMS or app).

This approach will enable a shift from a reactive logic to a predictive and proactive strategy, allowing early intervention in at-risk communities with adaptive, culturally sensitive, and co-designed tools.

The EVACS pilot could enable a continuous mechanism for evaluating and planning vaccination strategies by systematically analyzing endpoint results and integrating sentiment analysis as a feedback tool. This dynamic approach would allow real-time adjustments based on community response, improving both effectiveness and equity. Sentiment trends would serve as indicators of public perception, guiding the refinement of communication and engagement interventions.

References

- [1] Piano Nazionale per la Prevenzione Vaccinale 2023-2025. Ministero della Salute. https://www.salute.gov.it/imgs/C_17_pubblicazioni_3258_allegato.pdf
- [2] Piano Nazionale di Ripresa e Resilienza. Presidenza del Consiglio dei Ministri. <https://www.governo.it/sites/governo.it/files/PianoNazionaleRipresaResilienza.pdf>

- [3] Giacalone, M., Mazzola, V., & Di Nocera, F. (2021). Community participation and health promotion: the experience of a project in Sicily, Italy. *Annali di igiene: medicina preventiva e di comunità*, 33(4), 395-403.
- [4] Baldissera, S., Campostrini, S., & Perucci, C. A. (2019). How to measure vaccine coverage in Italy. *Annali di igiene: medicina preventiva e di comunità*, 31(5), 445-450.
- [5] Lorenzetti, S., Jones, M., & Brown, C. (2020). The role of telehealth during the COVID-19 pandemic across the interdisciplinary cancer team: Implications for practice. *Seminars in Oncology Nursing*, 36(5), 151082.
- [6] https://www.santannapisa.it/sites/default/files/community_building_oa_0.pdf
- [7] <https://www.degruyter.com/document/doi/10.1515/9781400841356.489/html>
- [8] <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0120701>

3.3. Digi-Vax: digitalization of vaccination processes and integration with surveillance systems - UNIPA

Introduction

Vaccine-preventable diseases (VPDs) are one of the main causes of morbidity, mortality, and healthcare costs, across all ages and specifically among newborns, children, elderly and frailty. The 2023–2025 Italian National Vaccine Prevention Plan (PNPV) represents one of the best model to prevent VPDs trough life course [1]. However, vaccination coverage rates remain lower than recommended in several Italian Regions, including Sicily that represents the fourth most populous with 4.8 million inhabitants [2].

Moreover, in comparison with other European and not-European Countries there is a lack of system that allow the evaluation of vaccine effectiveness, of vaccination coverage (with updated data), of vaccination safety and that generally allow an "easy to access" approach to vaccination for general population [3, 4, 5].

Mission 6 of Italy's National Recovery and Resilience Plan (PNRR) seeks to address this by prioritizing digital health transformation, including the ongoing development of a National Vaccine Registry.

DigiVax have the main objectives to:

1. make interoperable the data of Regional Vaccination Registry of Sicilian with the data of the Regional Reference Laboratory of Molecular diagnosis of VPDs (at the University Hospital "Paolo Giaccone" of Palermo);
2. make "easy to access" the vaccination process (reservation, consultation of vaccine status, adverse events reporting, etc...) and Interoperable with Personal Health Record and with existing App such as "IO".

Actually, was conducted a pilot on laboratory data of three vaccine preventable diseases such as Pneumococcal, Meningococcal and Haemophilus Influenzae (Invasive or non-Invasive bacterial diseases).

The DigiVax main aim is to demonstrate the feasibility and reproducibility of the project for other VPDs and for other Italian Regions.

Data Collection for Pilot project

The Sicilian Regional Vaccination registry allow to evaluate the vaccination status, date of vaccine administration, type of vaccine administered for all Sicilian General Population.

The Sicilian Reference Laboratory at the University Hospital of Palermo represents the reference for molecular diagnosis of all VPDs included in the Sicilian Vaccination Schedule such as Influenza, respiratory syncytial virus (VRS), COVID-19, measles, mumps, rubella, varicella, meningococcus, pneumococcus, haemophilus influenzae, etc...

All cases that were molecularly tested positive for three vaccine preventable diseases such as Pneumococcal, Meningococcal and Haemophilus Influenzae responsible for Invasive or non-Invasive bacterial diseases (meningitis, pneumonia, bacteremia) isolated from 2021 to march 2025 in Sicilian Hospital were cross-referenced with the data of the Sicilian Regional Vaccine Registry to assess vaccination history and identify, in case of vaccination, the type of vaccine administered, the date of the doses and if the vaccination schedule was completed or not.

Data collection and management (that were previously anonymized and codified in order for cross-reference the single case) occur via the Sicilian Vaccine registry, AVUR managed by Onit and the reference laboratory, managed by ModuLab, both GDPR-compliant platforms.

The cross-reference of data was manual and was related to vaccine available in Italy to prevent Pneumococcal, Meningococcal and Haemophilus Influenzae diseases. Overall, 33 cases of pneumococcal diseases, 8 of meningococcal diseases and 7 due to haemophilus spp. were analyzed and cross-referenced with vaccination data.

Technical details of DigiVax and its Applications

The DigiVax process need to develop algorithm that trough a software could help in make interoperable the two data (vaccine status and laboratory diagnosis) in order to analyze vaccine effectiveness of vaccination included in the 2023–2025 Italian National Vaccine Prevention Plan. Moreover, Digivax can help in modify and adapt vaccination strategies in case of approval of new vaccine formulation such as the new pneumococcal 21 valent vaccine in the prevention of Pneumococcal diseases in adult population [6].

As you can see in Figure 1, analyzing all cases of pnuemococcal invasive diseases in Sicilian region from 2021 to 2025 in subjects over 60 years old (that are the target of pneumococcal conjugate vaccines - PCV at 13, 15, 20, or 21 valence and pneumococcal polysaccharide vaccine - PPV at 23 valence) emerges that only the 21 valent have the possibility to protect 81% of cases, in comparison with PCV 20 or PPV 23 (the actual prevetive streategy in Sicily is one dose of PCV 20 follwed after 1 year by a dose of PPV 23, that protect in only 30% of cases).

It is Important also note that all cases are not vaccinated, and this confirm the need to Improve vaccination adherence especially among the elderly and fragile populations.

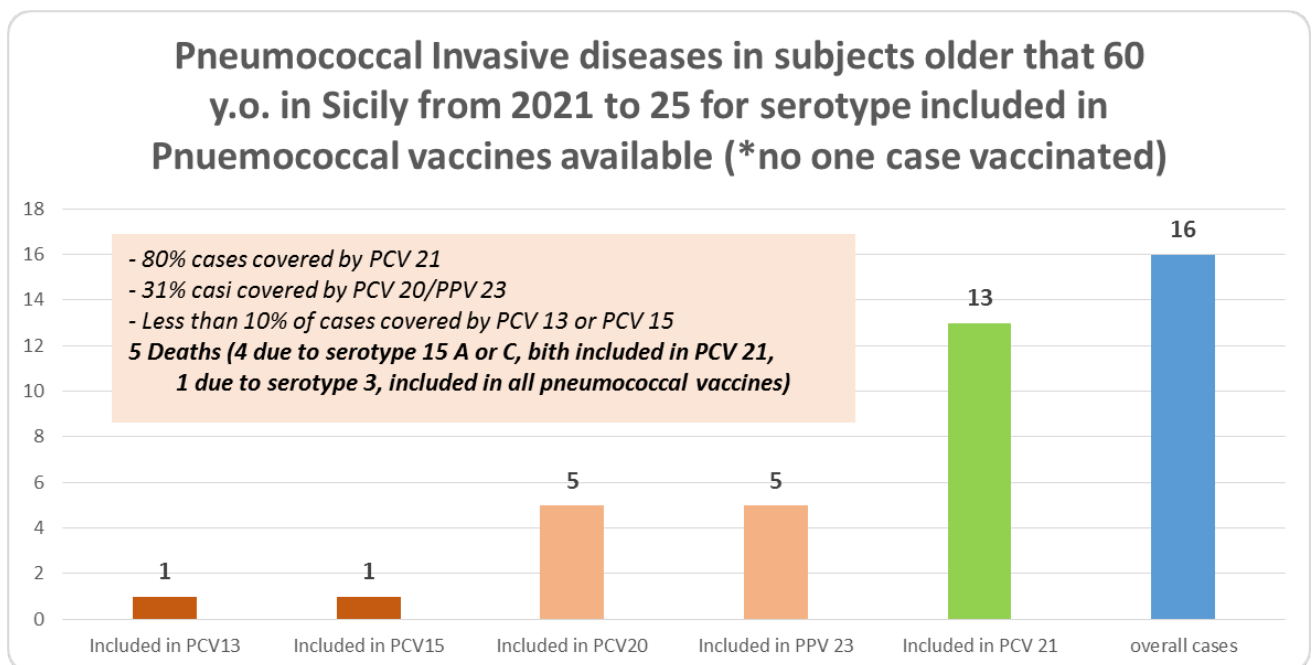


Figure 4. Pneumococcal cases among subjects aged over 60y.o. in Sicily from 2021 to 2025, related with vaccination status and protection of different pneumococcal vaccines authorized.

References

[1] Gazzetta Ufficiale. (2024): Ministero della Salute Piano Nazionale Prevenzione Vaccinale 2023-2025 Available at: <https://www.gazzettaufficiale.it/eli/id/2023/08/21/23A04685/sg>



[2] Istat, dati demografici. Available at: <https://www.demoistat.it>

[3] Andavac - Plan Vacunal estrategico de la Junta de Andalusia. Available at: <https://www.andavac.es/>

[4] Ausvaxsafety - Australia's active vaccine safety system. Available at: <https://www.ausvaxsafety.org.au/>

[5] CDC. ACIP Presentation Slides: April 15-16, 2025 Meeting. Available at: <https://www.cdc.gov/acip/meetings/presentation-slides-april-15-16-2025.html>

[6] Vaccinazione pneumococcica negli adulti, approvato in Italia V116. Available at: https://www.quotidianosanita.it/scienza-e-farmaci/articolo.php?articolo_id=129907

4. A sustainable and technological approach to large-scale prevention of falls and injuries

4.1. BONESTRENGTH: An *in-silico* clinical trial (ISCT) technology to assess the efficacy of intervention strategies for the prevention of hip fractures - UNIBO

Introduction

Osteoporosis (OP) is a major global health concern, affecting approximately 500 million people worldwide. The economic impact of OP is substantial, and as life expectancy continues to increase, particularly in developed countries, the burden of osteoporosis is expected to escalate, placing significant strain on healthcare systems [1]. Hip fragility fractures are one of the most severe consequences of OP, with around 20% mortality within a year and 60% of patients experiencing long-term disability. Therefore, the prevention of hip fracture is of major importance. Interventions to prevent hip fractures include lifestyle changes, pharmacological treatments, and fall prevention strategies.

BoneStrength is an In Silico trial technology originally developed for the assessment of osteoporosis treatments [2,3]. In this project, BoneStrength use is explored for primary prevention. The envisioned application is a platform based on BoneStrength, which could be used by prevention healthcare services to identify appropriate prevention strategies. Citizens above 55 years of age are identified as at general risk of fragility hip fracture; they need to be stratified into subpopulations for which certain prevention interventions can be cost-effective. The BoneStrength service could be used to predict the efficacy (in terms of reduced hip fracture incidence) of different prevention interventions in specific subpopulations.

Data Collection for Algorithm Development

The In Silico trial BoneStrength has been presented in previous studies [2,3]. Briefly, a methodology for generating virtual patients has been developed based on a morpho-densitometric atlas [4]. Finite Element (FE) models of each femur are used to predict subject-specific failure load, using a procedure previously validated against experimental measurements of failure load [5]. Subsequently, multiple side falls are simulated, with fall parameters stochastically sampled from a range of possible scenarios [6]. A patient is considered fractured when the impact force associated with a fall exceeds the femur strength, as shown in the Figure below. Multiple realisations of this stochastic Markov chain process are run to obtain convergence of predicted hip fracture incidence. Each realisation requires 2000-7000 FE simulations depending on the follow-up duration, which are solved using High-Performance Computing infrastructure (Leonardo, Cineca, Italy).

Technical Details of the Algorithm and its Applications

First, a novel approach to simulate fall events was integrated into the existing framework, to replicate the over-dispersed phenomenology of fall events in the population of interest. A Negative Binomial distribution was used to model the frequency of fall events in the population of interest, as reported in clinical data, where approximately 70% of individuals

are non-fallers, approximately 15% experience one fall, and 15% experience two or more falls [7]. Three virtual cohorts were simulated by replicating the control groups of three concluded clinical trials reported in the literature (LIFT, FREEDOM, ARCH) [8-10]. These were characterised by different fracture risk levels, to validate the ability of the model to predict fracture incidence in different populations. Subsequently, three intervention groups were simulated. A virtual cohort of 1594 patients was generated by match the distribution of baseline total hip areal bone mineral density (aBMD) as reported in a large-scale observational study [11]. The placebo group was simulated by updating the FE material properties over time, according to the aBMD changes reported in the reference study. In first approximation, the pharmacological treatment with bisphosphonates was simulated by assuming no further bone loss over time. A fall prevention strategy was simulated by adapting the Negative Binomial distribution parameters, assuming a 50% reduction in yearly fall rate for first-time fallers. Fracture incidence was predicted over a follow-up of 9 years and compared to assess the effectiveness of each intervention.

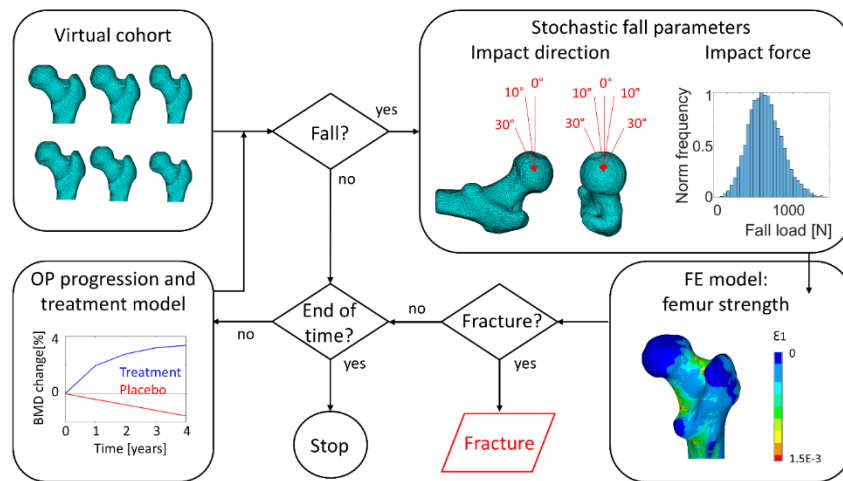


Figure 5. Conceptual framework of the In Silico trial BoneStrength.

For the LIFT cohort (N = 1270), the model predicted 12 ± 4 fractures over a three-year follow-up, compared to 8 fractures reported in the reference clinical trial. For the FREEDOM cohort (N = 1249) predicted fractures were 16 ± 3 in three years, with 14 observed in the clinical study. For the ARCH cohort (N = 1262), representing a high-risk population, the model predicted 37 ± 7 fractures over two years, compared to 41 fractures reported in the trial. In all cases, the predicted distribution consistently included the clinical data. The calibrated model showed an average 23% improvement in predictive accuracy compared to the previous implementation [3].

BoneStrength could predict a decrease in fracture incidence in the three intervention groups, as shown in the Figure below. Bisphosphonate treatment and fall prevention alone

achieved comparable effectiveness in fracture incidence reduction. The combination of pharmacological and fall prevention interventions achieved the highest efficacy, with a 43% reduction in fracture incidence compared to placebo.

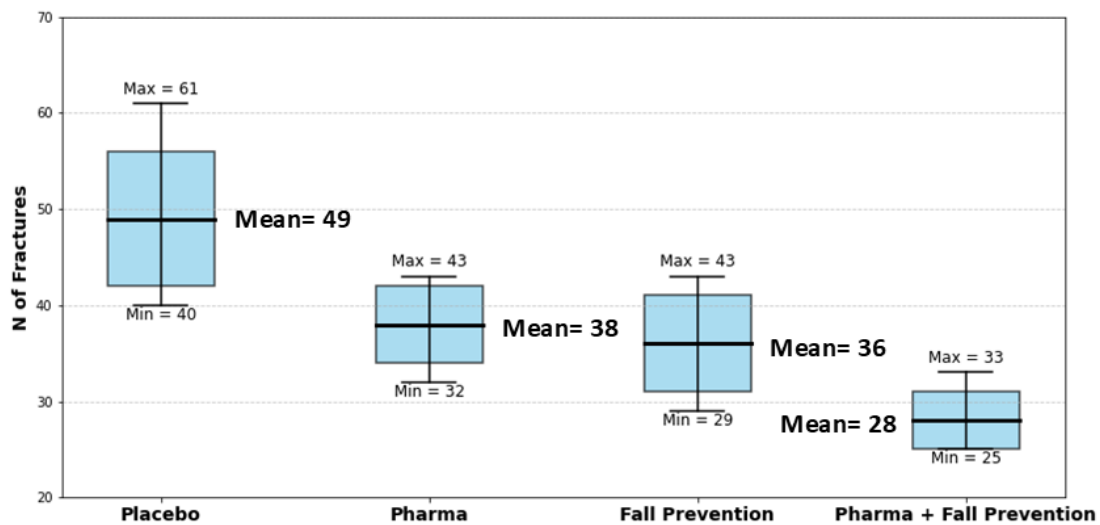


Figure 6. Predicted fracture incidence for the simulated placebo and intervention groups.

In conclusion, this In Silico trial methodology (BoneStrength) was able to predict fracture incidence in populations characterised by different risk profiles consistently with clinical data. In future developments, the effect of three different interventions (hip protectors, bisphosphonate treatment and fall prevention) will be simulated to predict and compare their effectiveness for reducing hip fracture incidence. This will constitute a proof-of-concept of a platform based on BoneStrength to identify an appropriate intervention for the prevention of hip fractures.

References

- [1] Wu, A.-M. et al. Global, regional, and national burden of bone fractures in 204 countries and territories, 1990–2019: a systematic analysis from the Global Burden of Disease Study 2019. *Lancet Healthy Long.* 2019; 2:e580-e592.
- [2] Oliviero, S., La Mattina, A. A., Savelli, G. and Viceconti, M. In Silico clinical trial to predict the efficacy of hip protectors for preventing hip fractures. *J Biomech.* 2024;176:112335.
- [3] Savelli, G., Oliviero, S., La Mattina, A. A. and Viceconti, M. In Silico Clinical Trial for Osteoporosis Treatments to Prevent Hip Fractures: Simulation of the Placebo Arm. *Ann Biomed Eng.* 2025;53:578-587.
- [4] La Mattina, A. A., Baruffaldi, F., Taylor, M. and Viceconti, M. Statistical Properties of a Virtual Cohort for In Silico Trials Generated with a Statistical Anatomy Atlas. *Ann Biomed Eng.* 2023;51:117-124.

- [5] Schileo, E., Balistreri, L., Grassi, L., Cristofolini, L. and Taddei, F. To what extent can linear finite element models of human femora predict failure under stance and fall loading configurations? *J Biomech.* 2014;47:3531-3538.
- [6] Bhattacharya, P., Altai, Z., Qasim, M. and Viceconti, M. A multiscale model to predict current absolute risk of femoral fracture in a postmenopausal population. *Biomech Model Mechanobiol.* 2009;18:301-318.
- [7] Ullah, S., Finch, C. F. and Day, L. Statistical modelling for falls count data. *Accident Anal Prev.* 2010; 42:384-392.
- [8] Cummings, S. R. et al. The effects of tibolone in older postmenopausal women. *N Engl J Med.* 2008;359:697-708.
- [9] Cummings, S. R. et al. Denosumab for Prevention of Fractures in Postmenopausal Women with Osteoporosis. *N Engl J Med.* 2009;361:756-765.
- [10] Saag, K. G. et al. Romosozumab or Alendronate for Fracture Prevention in Women with Osteoporosis. *N Engl J Med.* 2017;377:1417-1427.
- [11] Ensrud, K. E. et al. Frailty and Risk of Falls, Fracture, and Mortality in Older Women: The Study of Osteoporotic Fractures. *J Gerontol A.* 2007;62:744-751.

4.2. Muscle power and motor control degradation are better predictor of falls than muscle strength in the aging population - IOR and UNIBO

Introduction

Falls are a health risk and cause significant injuries (including fracture, subsequent surgical intervention, and postoperative course) which can further limit a person's ability to perform their daily activities. Studies suggest that in the presence of knee osteoarthritis (OA), the

risk of falling increases (+30%) [1]. OA is in fact associated with knee instability, and muscle weakness. While various biomechanical parameters, such as muscle strength, muscle power, motor control, quality and quantity of mobility, are known to contribute to (the risk of) falling, it is still unclear which one is the best predictor of a fall event. There are no studies who have comprehensively looked at all the above parameters in the context of fall prediction.

The PowerAging project was developed as a clinical protocol aimed to explore the relationship between age-related declines in muscle strength, muscle power, mobility, and motor control, and their association with fall risk. To address this, a longitudinal study involving older adults with knee osteoarthritis has been designed. The protocol includes repeated assessments over time using quantitative methods such as gait analysis, dynamometry, mobility monitoring through wearable sensors, medical imaging, and standardized clinical questionnaires.

Data Collection for Algorithm Development

This pilot study involves extensive data collection and a wide range of biomechanical and mobility-related parameters that will be explored and compared to identify predictors of functional decline that are expected to correlate with fall risk. Therefore, the target population comprises of 50 subjects with knee OA, aged 65 to 80 years old, who are at a higher risk of falling [1]. Participants will be monitored over a period of 18 months (for a total of 4 visits, every 6 months). All subjects will undergo a comprehensive assessment measuring (1) muscle power and muscle force, via isokinetic and isometric tests on a dynamometer, and a instrumented stairs ascent and descent test; (2) the subjects' motor function through a gait assessment; and (3) the level of mobility in the real world via the continuous monitoring over 5 days with a single waist-worn inertial sensor [2]. At first and last visits, a full lower limb MRI scan will be acquired to gather additional information on the muscular tissue. Clinical questionnaires and a bioimpedance analysis will complement the protocol.

Technical Details of the Algorithm and its Applications

The development of the predictive model begins with the elaboration and analysis of the collected experimental data. More specifically, the following steps will be carried out during the data processing phase:

- Muscle force will be assessed via isometric knee extension and flexion tests. Raw torque signals will be filtered and analysed to identify peak force and the initial rate of torque development.



- Muscle power will be evaluated through isokinetic testing, enabling the calculation of peak power, total work, and the power output during each extension or flexion movement.
- The instrumented stair ascent and descent test will be analysed to extract ground reaction forces and processed to establish the power to complete the motor task.
- Real-world mobility will be assessed through Digital Mobility Outcomes (DMOs), extracted using the Mobilise-D pipeline from the inertial sensor data continuously recorded over a 5-day period by a waist-worn sensor.

In addition, computer models and simulations will be employed to estimate biomechanical quantities of interest, such as muscle forces and joint contact forces, using traditional approaches (e.g., inverse approach) and testing novel tools to enable predictive simulations of human movement.

Upon data processing, a statistical framework will be applied to perform dimensionality reduction and variable selection, aiming to extract a limited set of variables that best summarize the information and capture the most clinically relevant aspects across domains and time points.

Potential correlations between data acquired across different domains – including real-world mobility (via a 5-day inertial sensor), muscle performance (via isometric and isokinetic tests), and internal biomechanical quantities (from musculoskeletal modelling) – will be investigated. Principal component analysis (PCA) may also be considered to reduce the number of variables, while preserving as much information as possible.

An alternative approach involves using the Joint and Individual Variation Explained (JIVE) [3] to integrate multiple domains (e.g., mobility, muscle force, and biomechanics), allowing the decomposition of joint and domain-specific components and the extraction of interpretable subject-specific scores. As the statistical approach must consider the longitudinal nature of the study and therefore of the inter-subject variability and subject-specific trajectories over time, adaptations of the above approach may be implemented. For example, following the methods proposed by [4], one could apply the JIVE to blocks of variables from the same or different domain, collected at different time points; thus enabling the projection of the original data onto shared directions of variation, yielding subject- and time-specific scores.

Regardless of the statistical approach, the selected relevant variables should describe the muscle tone and motor performance of the subject.

Once one or more relevant variables have been identified, they will be used to define a cost function to be integrated into a predictive simulation framework (e.g., within OpenSim Moco). This will enable the exploration of how changes in muscular capacity, joint loading,

or motor control strategies—such as those resulting from rehabilitation interventions—may affect mobility performance and stability, providing a powerful tool to simulate and test hypothetical interventions or disease progression scenarios *in silico*. Additionally, if some negative events occur, a threshold value could be proposed for a logistic regression to estimate the risk of falls. This model will constitute the basis for further activities to assess the sensitivity and specificity on a larger dataset.

References

- [1] Brand C, AW J, Lowe A, Morton C. et al. Prevalence, outcome and risk for falling in 155 ambulatory patients with rheumatic disease. *APLAR J Rheumatol*. 2005; 8:99–105.
- [2] Mikolaizak AS et al. Connecting real-world digital mobility assessment to clinical outcomes for regulatory and clinical endorsement—the Mobilise-D study protocol. *PloS One*. 2022; 17:e0269615.
- [3] Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat*. 2013; 7:523–542.
- [4] Jones MB, Merchant A, Morales-Soto L, Thompson JMD, Wall CR. The technique ‘joint and individual variance explained’ highlights persistent aspects of the diet using longitudinal food frequency data. *Br J Nutr*. 2022; 128:2054–2062.

4.3. DARE-FALLSPREDICT: development of a multi-variable model beyond the state of the art for estimating the risk of falling in older people - UNIBO and AUSL ROMAGNA

Introduction

The primary objective of the DARE-FALLSPREDICT project is to develop a model to estimate the one-year fall risk in older adults, with better performance than the FRAT-UP scale [1], which currently represents the state of the art. The aim is to improve the C-statistic of the FRAT-UP scale from 0.65 (estimated from the area under the ROC curve; [2]) to at least 0.75 after correction for optimism. Not only the discriminative ability of the model proposed here will be compared with that of the FRAT-UP score, but it will also be explored in comparison with simpler prediction strategies: these include a model in which previous falls are used to predict future ones and the screening algorithm proposed by World Guidelines [3].

The secondary objective is the development of a dynamic model to predict fall risk over six-month intervals. This will leverage the information acquired over time.

Exploratory objectives include the validation and calibration of the model in both frail and non-frail older adults and the estimation of the direct costs associated with falls, whether or not injuries are involved.

Data Collection for Algorithm Development

At time points T0 (baseline) and T1 (six months after baseline), data from the DARE-FALLSPREDICT study will be collected for both the computation of the FRAT-UP score and from sensors. The entire set of information collected at each time point will be referred to as X0 and X1, respectively. Fall data will instead be collected every month for twelve months. These will be aggregated into a binary indicator on a six-month window: Y0 will be equal to one if at least one fall occurs during the first six months following recruitment (and zero otherwise); similarly, Y1 will be equal to one if falls are recorded in the second semester. From this, the following temporal structure arises: X0 (explanatory variables at baseline) will be paired with Y0 (at least one fall in the first semester), while X1 (explanatory variables after six months) will be paired with Y1 (at least one fall in the second semester). The study was designed to recruit a total of 1,084 older adults, including 490 frail patients (Cohort 1) and 594 non-frail patients (Cohort 2). The sample size calculations [4] were performed with R using the pmsampsize package (available on <https://cran.r-project.org>) considering a shrinkage of predictor effects less than 10 %, a difference of 5% in the model apparent and adjusted Nagelkerke R2 value estimated from the C-statistic, an estimation within 5% of the average outcome risk in the population to address calibration issues, and correction for optimism.

For the purpose of developing a contingency plan to hedge against unexpected decreases in recruitment rate, sample size calculations have also been performed decreasing the number of predictors to be used in the prediction model. In particular, keeping the other factors that impact the criteria to be met unchanged according to the guidelines of [4], the new sample size will be 323 subjects for Cohort 1 and 264 subjects for Cohort 2. Cohort 1 will also include subjects recruited with a similar protocol from the general population in the DARE-FALLSPREDICT GP satellite pilot study, with an expected sample size of 200. The number of continuous predictors will be constrained between 8 and a minimum of 4 (contingency plan) through the techniques described below.

Technical Details of the Algorithm and its Applications

A two-step approach will be adopted:

1. Dimensionality reduction and variable selection. The GENEActiv sensor (DARE-FALLSPREDICT pilot study) or the Empatica EmbracePlus sensor (DARE-FALLSPREDICT GP satellite pilot study) will provide sleep quality measurements (amounting to more than ten variables), while the Dynaport7 MoveMonitor will

generate a similar number of variables related to gait quality in both pilot studies. Dimensionality reduction or variable selection techniques will be applied to these two domains to comply with the constraints on the number of predictors to include in the model. The objective is to extract, for each sensor, between 1 and 3 variables that can effectively summarize the information collected.

2. Definition of a prediction model. A multiple regression model will be defined to handle the nature of the response variable appropriately. The predictors will include the FRAT-UP score, along with either the components extracted or the variables selected during the previous phase.

In both phases, more than one methodology will be explored in order to make the most of the information collected in the study and to pursue the aforementioned objectives effectively.

The first strategy for dimensionality reduction is based on the method known as Joint and Individual Variance Explained (JIVE; [5]). JIVE is a direct extension of Principal Component Analysis (PCA) for data originating from multiple sources, and it allows the decomposition of data from multiple domains into the sum of three components: a low-rank approximation capturing the joint variation across domains, low-rank approximations capturing the individual variation within each domain, and residual noise. In the study by [6], JIVE is used as a dimensionality reduction technique designed to simultaneously analyze data from three domains: physical activity, sleep, and circadian rhythm.

Moreover, JIVE can be applied to blocks of variables from the same domain but collected at different time points [7]. It is possible to obtain scores for each time point and each individual by projecting the original data onto the directions of shared variability identified during estimation, according to the following formula:

$$S_t = X_t V_t$$

where:

- X_t contains the data at time t ,
- V_t represents the shared variability directions,
- S_t denotes scores for all individuals at time t .

These projected scores will serve as predictors in the risk prediction model. The ability of these scores to capture domain-specific variability will be evaluated using techniques described in [7]. JIVE can be easily implemented in R using the `r.jive` package (CRAN: Package `r.jive`).

Alternative strategies include performing a preliminary selection of the longitudinal variables that contribute most to predicting the outcome of interest. These selected variables can then be fed to a mixed-effects logistic regression model that accounts for the longitudinal nature of the data. To this end, variable selection algorithms such as SES

(Statistically Equivalent Signatures) may be used. SES aims to identify the smallest subset(s) of predictors with the highest predictive performance for a given target variable. In particular, we will use the version adapted for longitudinal data, available via the `ses.temporal` function from the `MXM` package [8]. In order to avoid inflated model accuracy estimates and suboptimal feature selection, we will follow the guidelines outlined in the tutorial by [9]. Furthermore, the absence of multicollinearity issues will be checked using Variance Inflation Factors (VIF).

Two distinct model specifications are required in order to:

1. obtain an overall estimate of fall risk across the entire follow-up period;
2. dynamically model the risk of falls over six-month time windows.

In the first case, i.e., model prediction of the risk of fall over one year of follow-up:

- the response variable is defined as the maximum of Y_0 and Y_1 ;
- summary indicators for the sleep and gait domains can be derived either from JIVE or from other techniques that synthesize only the information contained in X_0 .

Given the binary nature of the response variable — that takes value 1 if a fall occurs and 0 otherwise — it is possible to estimate fall risk using logistic regression. Predictors will include the FRAT-UP score achieved by each subject, as well as summary indicators for the sleep and gait domains. The latter will consist of both the joint and individual scores returned by JIVE.

This static model can be adapted to predict the risk of fall in any subset of the follow-up period (from 1 to 11 months), assessing the impact of this modelling choice on predictive power.

For the second case, it is important to account for the information accrued in the different available time points. Following [7], JIVE projections at time T_1 (S_1) will be exploited, together with the FRAT-UP score, to predict the risk of fall in the second semester (Y_1). If we are successful in extending the study, this procedure will yield predictions for the risk of fall in any additional semester – e.g., S_2 will be used to predict Y_2 , etc.

Alternative strategies for dynamic modelling include logistic regression with mixed effects. This accounts for the fact that repeated measurements of the same variable for the same individual over time tend to be more similar than measurements taken from different individuals. A generalized linear mixed model will be fitted, where:

- at the first time point, X_0 is paired with Y_0 , while at the second time point, X_1 is paired with Y_1 ;
- for the sleep and gait domains, `ses.temporal` will be applied for variable selection. Alternatively, it will be assumed that the variables selected based on the information available at the first time point remain the most relevant for the second.

- a random effect for the individual will also be included, with the simplest assumption being a subject-specific intercept;
- the FRAT-UP score will be used as predictor also in this model.

Two of the most used R packages for estimating mixed-effects models are nlme and lme4 (CRAN:<https://cran.r-project.org/web/packages/lme4/index.html>).

In addition to the investigation described above, further exploratory analyses will be performed using information from the EmbracePlus sensor for cardiac activity worn by the subjects recruited as part of the DARE-FALLSPREDICT GP study.

The model will be evaluated in terms of its discriminative ability — that is, how well it distinguishes between patients who experienced a fall event and those who did not. This aspect can also be interpreted as the ability of the model to classify patients into low or high-risk categories for falls. The Receiver Operating Characteristic (ROC) curve analysis is traditionally used to assess this capability. The area under the ROC curve (AUC), also known as the C-statistic, provides a summary measure of the model's prognostic performance and can be easily implemented using the pROC package (available on <https://cran.r-project.org>). When developing a prediction model, the available data can be used both for model fitting and model assessment. However, in this case, performance is said to be apparent or re-substituted, and its estimates can be optimistic [10]. For the specific case of binary outcomes (as is the case of this study), the literature contains several papers comparing different methods for correcting the optimism of the apparent area under the ROC curve (e.g., [11]). Some of the techniques that have shown greater effectiveness in handling optimism correction and that could, therefore, be employed in the present analysis include k-fold cross-validation with replication and bootstrap validation. These methods can be implemented in R using packages such as caret, boot, and rms, which provide flexible tools for resampling, model validation, and optimism adjustment. All packages are available on CRAN (<https://cran.r-project.org>).

Model calibration will be of great importance, as poor calibration can make an algorithm less useful in clinical practice than another model with a lower AUC but better calibration [12]. Calibration evaluates the agreement between predicted probabilities and the actual proportions of observed outcomes and is an essential item in the TRIPOD and TRIPOD-AI guidelines [13]. One way to examine the calibration of risk predictions is through calibration curves. These will be generated using CalibrationCurves package in R (<https://cran.r-project.org/web/packages/CalibrationCurves/index.html>). Summary calibration indicators, such as the calibration-in-the-large and calibration slope, will also be reported to quantify systematic miscalibration.

Multiple methods will be applied to improve model calibration. One such approach is Platt Scaling, namely a univariate logistic regression that uses the model predictions as

independent variables and the binary outcomes as the dependent variable. In R, this can be implemented using the `glm` function with a logit link. Other techniques will also be explored depending on model performance and complexity.

The study lends itself to several exploratory analyses. A first example concerns the analysis of fall risk in two distinct cohorts: a low-risk group (Cohort 2, for which the literature estimates a fall risk of 22%) and a high-risk group (Cohort 1, for which the literature estimates a considerably higher fall risk equal to 30%). The study includes an interview to assess frailty using the Clinical Frailty Scale (CFS), which determines the cohort assignment. The main objective in this context is to compare the observed fall risk in the two cohorts with the values reported in the literature in order to assess consistency or identify potential discrepancies.

Data collected in the DARE-FALLSPEDICT study during motor tests, such as the Timed Up and Go (TUG) test and the Romberg test, will also be used to construct alternative exploratory predictors with greater discriminative power.

Moreover, the availability of temporal data allows for the analysis of fall frequency throughout the year, with the aim of identifying potential recurring or seasonal patterns. For this purpose, advanced statistical tools such as the `timeOmics` package [14] will be employed, as it is particularly suitable for seasonality analysis in biomedical settings.

An additional exploratory approach will involve achieving the objectives through survival analysis, which studies the time until the occurrence of an event of interest. Specifically, three Cox-based models for recurrent event data will be employed: Andersen-Gill, Prentice-Williams-Petersen, and Wei-Lin-Weissfeld.

One could investigate the predictive ability of models that do not include FRAT-up and that do not require at least one component or variable per domain. In this context, the Sparse Group Lasso (SGL) can be effectively employed to achieve this goal. SGL is a regularization technique that combines L1 (lasso) penalization to induce sparsity at the level of individual coefficients, with L2 (group lasso) penalization to induce sparsity at the group level. The Sparse Group Lasso is therefore capable of selecting which groups of predictors are important and, within those selected groups, identifying which individual predictors are significant. The `sparsegl` package is available on CRAN.

Another area of significant interest will be the evaluation of the impact of the prediction model using the methodological framework proposed by [15], specifically in relation to observational studies.

The clinical usefulness of the prediction model will also be further investigated. Beyond traditional performance metrics, evaluating clinical utility means considering the real-world implications of using the model in practice—such as its impact on patient outcomes, feasibility of implementation, and cost-effectiveness. In this regard, methods such as

Decision Curve Analysis (DCA)—for instance, through the *dcurves* package in R—offer a robust framework for quantifying the net benefit of a model across different decision thresholds, thereby informing its value in guiding clinical decisions.

Lastly, although a recent systematic review found no clear performance advantage of machine learning over logistic regression for clinical prediction models [16], comparing our proposed models with selected ML algorithms may still prove valuable. Among the most promising ML approaches — and already applied in fall risk prediction [17] — is Gradient Boosting, implemented in the *gbm* R package. Evaluating its applicability to our dataset, with appropriate hyperparameter tuning and comparing its performance with logistic regression will be of particular interest.

We will also consider more interpretable variants of Gradient Boosting, such as Componentwise Gradient Boosting, implemented in the *mboost* package and available on CRAN (<https://cran.r-project.org>).

As for the time-dependent model to predict fall risk over six-month intervals, we will employ Mixed-Effects Machine Learning (MEML) approaches, such as the *GLMMtree* [18], available in the *glmertree* R package. As noted by the authors, this flexible decision tree algorithm provides a valuable data-analytic tool for clinical prediction tasks.

The roadmap presented in the Figure below outlines the main analytical phases planned in the project. The timelines indicated are to be considered approximate and may be revised based on the actual progress of recruitment and data availability.

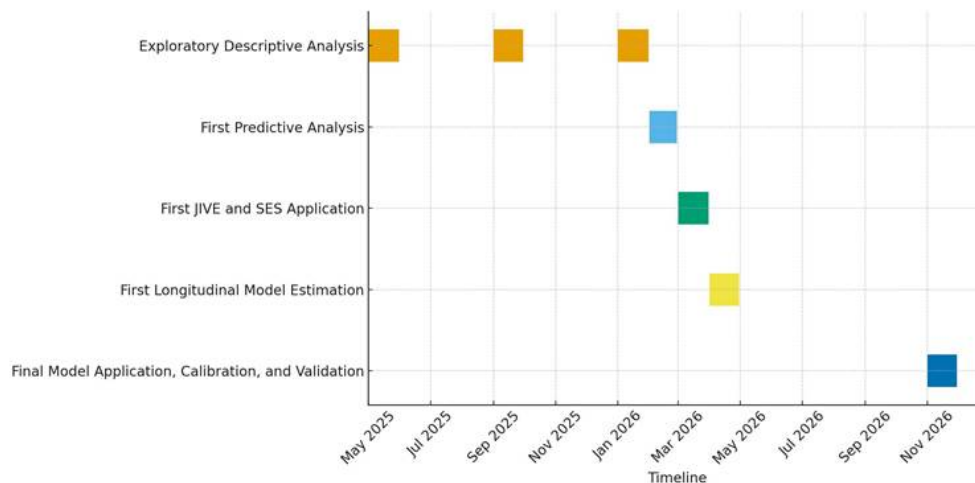


Figure 7. Proposed Roadmap.

References

[1] Cattelani L, Palumbo P, Palmerini L, Bandinelli S, Chiari L. FRAT-up, a Web-based fall-risk assessment tool for elderly people living in the community. *J Med Internet Res.* 2015;17:e41.

- [2] Palumbo P et al. Predictive performance of a fall risk assessment tool for community-dwelling older people (FRAT-up) in 4 European cohorts. *J Am Med Dir Assoc.* 2016;17:1106–1113.
- [3] Montero-Odasso M et al. Task Force on Global Guidelines for Falls in Older Adults. World guidelines for falls prevention and management for older adults: a global initiative. *Age Ageing.* 2022;51:afac205.
- [4] Riley RD, Snell KIE, Ensor J, Burke DL, Harrell FE, Moons KGM, Collins GS. Minimum sample size for developing a multivariable prediction model: Part II – binary and time-to-event outcomes. *Stat Med.* 2019;38:1276–1296.
- [5] Lock EF, Hoadley KA, Marron JS, Nobel AB. Joint and individual variation explained (JIVE) for integrated analysis of multiple data types. *Ann Appl Stat.* 2013;7:523–542.
- [6] Di J et al. Joint and Individual Representation of Domains of Physical Activity, Sleep, and Circadian Rhythmicity. *Stat Biosci.* 2019;11:371-402.
- [7] Jones MB, Merchant A, Morales-Soto L, Thompson JMD, Wall CR. The technique ‘joint and individual variance explained’ highlights persistent aspects of the diet using longitudinal food frequency data. *Br J Nutr.* 2022;128:2054–2062.
- [8] Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., & Tsamardinos, I. Feature Selection with the R Package MXM: Discovering Statistically Equivalent Feature Subsets. *J Stat Soft.* 2017; 80:1–25.
- [9] Huang S. Supervised feature selection: A tutorial. *Artif Intell Res.* 2015;10.5430:air.v4n2p22.
- [10] Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. *J Am Stat Assoc.* 1983;78:316–331.
- [11] Iparragirre A, Barrio I, Rodríguez-Alvarez MX. On the optimism correction of the area under the receiver operating characteristic curve in logistic prediction models. *SORT.* 2019;43:145–162.
- [12] Van Calster B, McLernon DJ, van Smeden M, Wynants L, Steyerberg EW. Calibration: the Achilles heel of predictive analytics. *BMC Med.* 2019;17:230.
- [13] Collins GS et al. TRIPOD+AI statement: updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ.* 2024;385:e078378.
- [14] Bodein A, Scott-Boyer M-P, Perin O, Le Cao K-A, Droit A. timeOmics: an R package for longitudinal multi-omics data integration. *Bioinformatics.* 2022; 38:577-579.
- [15] Palumbo P. Qini curves for potential impact assessment of risk predictive models informing intervention policies. *Value Health.* 2025. ISSN: 1098-3015.
- [16] Jie M. et al. A systematic review shows no performance benefit of machine learning over logistic regression for clinical prediction models. *J Clin Epidemiol.* 2019;110:12-22.

[17] Jahangiri S., Abdollahi M., Patil R., Rashedi E., Azadeh-Fard N., An inpatient fall risk assessment tool: Application of machine learning models on intrinsic and extrinsic risk factors, *Machine Learn Appl.* 2024;100519,2666-8270.

[18] Fokkema, M., Edbrooke-Childs, J., & Wolpert, M. Generalized linear mixed-model (GLMM) trees: A flexible decision-tree method for multilevel and longitudinal data. *Psychother Res.* 2020;31:329–341.

5. Innovative digital tools for personalized cardiovascular primary prevention - UCSC

Summary

In recent years, Polygenic Risk Scores (PRS) have emerged as crucial tools in the early prediction of diseases, revolutionizing the way we approach human health. These scores, which take into account the complex interplay of multiple genetic variants, make it possible to identify individuals at increased risk of developing various conditions—particularly in the fields of cardiovascular disease and cancer. This approach not only offers unprecedented opportunities for targeted preventive interventions but also holds the potential to transform clinical management, significantly improving health outcomes at both individual and population levels.

Based on this principle, within the DARE project, efforts have been made to develop pilot initiatives aimed at creating more precise and effective predictive algorithms to identify individuals at increased risk of certain chronic diseases.

The following four studies integrate the PRS into various clinical studies to enhance the prediction and prevention of cancer and cardiovascular diseases.

One pre-post trial focuses on healthy individuals receiving personalized lifestyle advice and PRS-based genetic risk evaluation for cardiovascular diseases, with baseline and final measurements compared to assess behavioural changes.

In parallel, a case-control study investigates the association between PRS and epithelial ovarian cancer, comparing affected women with healthy controls. Clinical, genetic, and lifestyle data are collected using electronic case report forms, and DNA analysis is performed to evaluate PRS and key mutations (e.g., BRCA1/2). Logistic regression models will assess the predictive power of PRS.

Another study explores PRS integration into CanRisk assessments for breast cancer. Women undergoing genetic screening provide feedback on their experience, while healthcare providers evaluate the feasibility and utility of PRS in clinical workflows.

Finally, a study on pancreatic cancer (PDAC) examines the relationship between PRS and disease risk, as well as its integration with known risk factors like smoking and diabetes. Associations between PRS, pathological features, and treatment strategies are analyzed to develop a multifactorial risk model.

Across all studies, PRS is used as a powerful tool to refine risk stratification, guide prevention, and potentially personalize clinical decision-making.

Polygenic Risk Score

The Polygenic Risk Score (PRS) is a quantitative measure of genetic susceptibility, derived by aggregating the effects of multiple single nucleotide polymorphisms (SNPs) associated with a specific trait or disease. These SNPs are identified through genome-wide association studies (GWAS), with each assigned a weight corresponding to its estimated effect size. Although individual SNPs contribute modestly to genetic risk, their combined effect can provide a more robust association with the disease. This aggregation offers a comprehensive assessment of genetic predisposition by capturing the cumulative burden of multiple risk variants. For this analysis, SNPs will be selected based on the most recent GWAS findings related to cancers and cardiovascular disease (CVD). The PRS will be computed as a weighted sum of risk alleles, with the weights reflecting the effect sizes reported in the literature. We will prioritize PRS that have demonstrated strong predictive performance in previous studies and validate them in our independent cohort to assess their generalizability and robustness across different populations.

The following section presents the four pilot studies, a brief introduction, the data collection for algorithm development process, and the technical details of the algorithm and its application.

5.1. Personalised HeartCare (PHC): innovative approaches for personalized primary prevention of cardiovascular diseases (CVDs)

Introduction

This pilot aims to evaluate if lifestyle changes occur following communication of the Polygenic Risk Score (PRS), using a validated questionnaire, and to evaluate the feasibility of incorporating PRS into patient care pathways.

In this pre-post-trial, we will compare the participants' measurements taken at baseline with those obtained at the end of the study. Each participant will serve as their own control, and we will assess whether any changes occur as a result of our intervention. We will enrol healthy participants without established cardiovascular disease, diabetes, or familial hypercholesterolemia, with no age restrictions. All differences will be considered during the analysis and results phases. The baseline assessment will involve blood testing and physical examination. Participants will then complete a lifestyle questionnaire, which assesses factors such as smoking status, alcohol consumption, dietary patterns, sleep patterns, and physical activity. Participants will be classified into three categories: favourable, intermediate, and unfavourable. These baseline measurements will serve as the reference for our study. Subsequently, each participant will receive personalized advice for his known risk factors. Furthermore, a blood sample will be taken from each participant, in order to collect DNA. Then the DNA will be analysed, and the genetic risk of developing CVDs using the Polygenic Risk Score (PRS), will be calculated. At the end of the trial, we will compare the results of each arm with the baseline measurements and, eventually, between the two arms, to assess whether the proposed intervention have resulted in lifestyle modifications.

Data Collection for Algorithm Development

The baseline assessment will involve blood testing and physical examination. Participants will then complete a lifestyle questionnaire, which assesses factors such as smoking status, alcohol consumption, dietary patterns, sleep patterns, and physical activity. Furthermore, a blood sample will be taken from each participant, in order to collect DNA. Then the DNA will be analysed, and the genetic risk of developing CVDs using the Polygenic Risk Score (PRS), will be calculated.

Technical Details of the Algorithm and Its Application

Polygenic Risk Score algorithm, quality control and validation

Quality control (QC) of samples and genotyping data will be performed using the Axiom™ Analysis Suite to ensure robust SNP filtering and the application of advanced genotyping methods prior to downstream analysis. We will exclude SNPs with a minor allele frequency (MAF) < 1%, SNPs that fail the Hardy–Weinberg equilibrium (HWE) test with $p < 10^{-6}$ for the total sample and $p < 10^{-10}$ for controls and cases, respectively. Furthermore, SNPs and samples with a call rate below 90% will be excluded to ensure data quality and reliability. To address potential sources of bias, including relatedness, sex discrepancies, and ancestry differences, additional QC measures will be applied. To minimize the influence of relatedness, we will assess genetic relatedness using identity-by-descent (IBD) analysis, excluding pairs of individuals with high genetic relatedness (e.g., second-degree relatives or closer). Sex discrepancies will be addressed by cross-verifying the reported sex of participants with genetically inferred sex based on sex-specific markers (e.g., X and Y chromosome variants), excluding samples with discrepancies between self-reported and genetically inferred sex. To account for population stratification, particularly with respect to European ancestry, principal component analysis (PCA) will be performed using the 1000 Genomes reference panel to identify and correct for ancestry-related differences. Individuals whose genetic background deviates significantly from European ancestry will be excluded to ensure a genetically homogeneous cohort. These QC measures will help minimize confounding factors and improve the validity and generalizability of the findings. Following the initial QC, imputation will be conducted using the Haplotype Reference Consortium (HRC) in the Helmholtz Munich Imputation Server to infer missing genotypes. This process aims to enhance the coverage of genetic variants and increase the power of subsequent analyses. Post-imputation QC will be applied to ensure the reliability and quality of the imputed genotypes, with variants exhibiting low imputation quality scores, high genotype uncertainty, or deviations from Hardy-Weinberg equilibrium being excluded. Additionally, SNPs with low MAF will be filtered out due to their higher error rates and limited contribution to the analysis. All QC procedures will be conducted using PLINK. After QC and imputation, Polygenic Risk Scores (PRS) will be standardized across the cohort. Individuals will then be stratified into predefined risk categories using percentile thresholds, enabling the identification of risk gradients within the study population.

The traditional model will incorporate well-established risk factors, including demographic variables (e.g., age, sex), clinical history (e.g., family history of disease, comorbidities), and lifestyle variables (e.g., smoking status, BMI, physical activity). The PRS will be derived from validated GWAS and tailored to the specific disease outcomes. We will construct and compare two sets of models for each condition: one based solely on conventional risk factors, and another incorporating both traditional predictors and the PRS. Model

performance will be evaluated using key statistical metrics, including discriminative ability (area under the receiver operating characteristic curve [AUC]), calibration (e.g., calibration plots, Hosmer-Lemeshow test), and reclassification improvement (e.g., net reclassification improvement [NRI] and integrated discrimination improvement [IDI]). Internal validation using cross-validation and bootstrap resampling techniques will be conducted to assess model robustness. This comprehensive modeling framework will allow us to quantify the incremental value of PRS in enhancing individual risk prediction for both cancer and CVD, and inform their potential integration into personalized prevention and screening strategies.

For this study, the PRS used is an existing, validated PRS based on a European population, from the study by Khera et al (1). The Polygenic Risk Score (PRS) developed by Khera et al. includes 6,630,150 single nucleotide polymorphisms (SNPs) associated with coronary artery disease (CAD). These SNPs were selected based on genome-wide association studies (GWAS) and weighted according to their effect sizes to estimate an individual's genetic risk for cardiovascular disease.

Variables collected by the pilot

This pilot will collect variables on the lifestyle of the participants, using the Life's Essential 8 (LE8). The LE8 is a comprehensive cardiovascular health assessment tool developed by the American Heart Association (AHA) to evaluate an individual's lifestyle and health factors. It includes eight key metrics, categorized into behavioral (diet, physical activity, nicotine exposure, and sleep health) and health-related (BMI, blood glucose, cholesterol, and blood pressure) components. Each metric is scored as follows:

- Diet quality: based on adherence to a heart-healthy eating pattern.
- Physical activity: minutes per week of moderate-to-vigorous activity.
- Nicotine exposure: status as a current, former, or never smoker, including vaping.
- Sleep duration: ideal range of 7–9 hours per night.
- BMI: scored based on optimal weight categories.
- Blood glucose, cholesterol and blood pressure: scored according to clinical guidelines for ideal levels.

Furthermore, we will collect data on the socioeconomic status, demographic information, knowledge of PRS and cardiovascular disease, level of anxiety, and reaction to the test of the participants. Finally, to explore the acceptability and feasibility of the intervention, an ad hoc questionnaire will be administered.

Each of the lifestyle variables collected by the Life's Essential 8 Each is scored on a continuous scale from 0 to 100, with higher scores indicating better cardiovascular health. Based on the final score, individuals are classified into three categories: poor (0–49),

intermediate (50–79), or ideal (80–100) cardiovascular health. The calculation will be integrated into REDCap.

Participants will then receive an integrated health assessment based on their Life's Essential 8 (LE8) score and Polygenic Risk Score (PRS). The LE8 score reflects modifiable lifestyle factors, while the PRS indicates genetic susceptibility to cardiovascular disease. By combining these two measures, a more tailored and precise set of recommendations can be provided. Individuals with high genetic risk may be encouraged to adopt stricter preventive measures, even if their lifestyle is currently favourable. Conversely, those with poor lifestyle habits but low genetic risk will still receive guidance to improve their cardiovascular health. Baseline characteristics of the four arms will be summarized using descriptive statistics.

The baseline values will be compared with the final value to evaluate the effect of the communication of the Polygenic Risk Score. The analysis will employ adjusted mixed-effect models for repeated measures to assess significant differences in lifestyle patterns, lipid, and CVD risk profiles from baseline to the follow-up time point (T).

We will conduct a comprehensive subgroup analysis to explore potential moderating and mediating effects based on participant characteristics. The subgroup analysis will encompass sociodemographic characteristics (age above or below 55 years, gender, and marital status), ethnicity, socioeconomic status (occupation and related variables), education level, PRS levels (high, intermediate, normal), lifestyle category at baseline (favorable, intermediate, or unfavorable), knowledge toward CVDs (present/absent), and psychological status at baseline. All statistical analyses will be conducted using STATA (StataCorp, USA) and R (<https://www.r-project.org/>).

5.2. Evaluation of Polygenic Risk Score for epithelial OVarian cancer risk prediction and clinical outcomes in an Italian population: the PROVE study.

Introduction

In this prospective case-control study, we will compare PRS evaluation between women with histologically proven epithelial ovarian and fallopian tube cancer (cases) and women with no personal history of epithelial ovarian cancer (controls). The aim is to evaluate the association of Polygenic Risk Score (PRS) and epithelial ovarian cancer risk. We will enroll participants aged >18 yrs with no concomitant malignancies, except ovarian cancer (OC) for cases, and no personal history of OC or self-reported previous bilateral oophorectomy for controls. At enrollment, women will be asked to complete a questionnaire regarding socio-economic characteristics, lifestyle and diet habits, either through REDCap platform or a paper form. Questions will be administered in native language. Moreover, to perform study-specific molecular analysis and biobanking, 2 ml and 6 ml of blood will be drawn from cases and controls respectively. An aliquot of 2 ml will be processed for the PRS analysis using the GeneTitan™ MC Fast Scan Instrument (220V). PRS results will be provided to all the participants, upon request.

For cases, genetic tests validated for diagnosis will be performed according to clinical practice and shared with the participants by the gynaecology oncologist or geneticist as appropriate. For controls, *BRCA1-2* tests validated for diagnosis will be performed and, if a *BRCA1-2* pathogenetic mutation will be found, a genetic consult will be held. Each participant will undergo a blood test, from which DNA will be extracted for *BRCA1-2*, *PALB2*, *RAD51C*, *RAD51D* pathogenetic variants and PRS evaluation by a Genome Wide Association Study (GWAS) approach. Logistic regression will be used to examine the association between the PRS and the outcome (OC yes/no). PRS will be calculated as the sum of an individual's risk alleles, weighted by risk allele effect sizes derived from GWAS and PRS will be grouped in different percentiles. Multivariable logistic regression will be applied to evaluate the adjusted role of PRS in ovarian cancer onset prediction.

Data Collection for Algorithm Development

At enrollment, women will be asked to complete a questionnaire regarding socio-economic characteristics, lifestyle and diet habits, either through REDCap platform or a paper form.

For the study, a customized electronic Case Report Form (eCRF) on REDCap platform will be used to collect clinical data of enrolled patients by pseudo-anonymization. Questions will be administered in native language. Moreover, to perform study-specific molecular analysis and biobanking, 2 ml and 6 ml of blood will be drawn from cases and controls respectively. An aliquot of 2 ml will be processed for the PRS analysis using the GeneTitan™ MC Fast Scan Instrument (220V). PRS results will be provided to all the participants, upon request.

Technical Details of the Algorithm and Its Application

All the information regarding quality control and validation of the PRS algorithm could be found above (please see chapter “*Polygenic Risk Score algorithm, quality control and validation*”).

All enrolled women will undergo *BRCA1-2*, *PALB2*, *RAD51C*, *RAD51D* pathogenetic variants detection in addition to the evaluation of their polygenic risk score (PRS) using the GeneTitan™ MC Fast Scan Instrument (international, 220V) and the Axiom™ Precision Medicine Diversity Array Plus Kit, 96-format (PMDA). The minimum required DNA quantity required for this analysis is 100–200 ng.

For the PRS analysis, SNPs for genotyping will be selected based on the latest findings from GWAS on EOC, particularly leveraging the results, as reported by Dareng et al (2). The PRS will be calculated as a weighted sum of risk alleles based on the selected SNPs. Each SNP will be assigned a weight according to its effect size, as reported in GWAS that have identified significant associations with EOC risk. These effect sizes quantify the relative contribution of each SNP to the overall genetic predisposition to the disease (3)(4)(5).

Moreover, all controls will undergo testing for *BRCA1* and *BRCA2* using a method that has been validated for diagnostic purposes. This analysis will be performed utilizing Devyser's BRCA next-generation sequencing kit (Devyser, Hägersten, Sweden). Following DNA extraction, the entire coding region of both *BRCA1* and *BRCA2* genes, including 10 to 20 base pairs of intronic flanking sequences around all coding exons, will be amplified using Devyser's BRCA next-generation sequencing kit (manufactured by Devyser, Hägersten, Sweden). Subsequently, sequencing will be performed using the Illumina MiSeq system (Illumina, San Diego, CA, USA). Raw sequence data analysis including base calling, demultiplexing, alignment to the hg19 human reference genome (Genome Reference Consortium GRCh37), and variant calling will be conducted using the CE-IVD Amplicon Suite Software (SmartSeq, Novara, Italy). This process will detect Single Nucleotide

Variations (SNVs), indels, and perform bioinformatics-based Copy Number Variation (CNV) predictions. If participants have previously undergone validated *BRCA1* or *BRCA2* testing for diagnostic purposes, these tests will not be repeated.

Sample size was calculated according to the results reported by Yang et al 2018 who evaluated the PRS for ovarian cancer risk prediction in a prospective cohort study [6] and managing PRS as a categorical variable (percentile categories). Samples of 650 subjects in the control group and 650 subjects in the case group achieve 90% power to detect a difference between the PRS proportions of the controls and the PRS proportions of the cases when the significance level (alpha) is 0.05. The PRS proportions in the two groups used for the calculation were taken from the previously mentioned Yang study. The sample size was calculated using the software PASS (PASS 2021 Power Analysis and Sample Size Software (2021). NCSS, LLC. Kaysville, Utah, USA) and the analysis "Tests for Two Ordered Categorical Variables (Non-Proportional Odds).

Patients' characteristics will be described as absolute frequencies and percentages for nominal variables and as medians and ranges or means and standard deviations for continuous variables, as appropriate. The normality of continuous variables will be assessed with the Shapiro-Francia test. Logistic regression will be used to examine the association between the PRS and the outcome (OC yes/no). PRS will be calculated as the sum of an individual's risk alleles, weighted by risk allele effect sizes derived from genome-wide association study data, and it will be evaluated both as a continuous variable and a categorical one. The PRS will be grouped into the percentiles: [0,5%), [5%,10%), [10%,20%), [20%,40%), [40%,60%), [60%,80%), [80%,90%), [90%,95%), and [95%,100%] based on the PRS distribution in controls, with [0,5%) as the lowest 5% PRS group and [95%,100%] as the highest. The middle [40%,60%) group will be used as the reference category. Moreover, multivariable logistic regression (adjusted by age, BMI, age at menarche, parity, use of oral contraceptive, menopausal status, hormonal therapy replacement for menopause and *BRCA* status) will be applied to evaluate the adjusted role of PRS in ovarian cancer onset prediction. All estimates will be presented with two-sided 95% Confidence Intervals (CIs). All reported p values will be two-sided, and a value of less than 0.05 will be considered statistically significant. No imputation will be carried out for missing data. Statistical analysis will be performed using STATA software (STATA/BE 17.0 for Windows, StataCorp LP, College Station, TX 77845, USA).

For eligible women the following data will be collected: socio-demographic characteristics, level of education according to International Standard Classification of Education (ISCED), type of employment, age at enrolment, age at menarche, body Mass Index (BMI), smoke and diet habits, physical activity, diagnosis of endometriosis, parity, age at last pregnancy,

breast feeding, use of oral contraceptive, tubal ligation, menopausal status, age at menopause (if applicable), use of hormonal replacement therapy (HRT) in menopause, personal and family history of cancer (in particular ovarian; endometrial and gastrointestinal), *BRCA1*, *BRCA2*, *PALB2*, *RAD51C*, *RAD51D* pathogenetic variants and PRS.

5.3. Integrated Genetic Risk Models (MIG) with Digital Solutions to Transform Breast Cancer Prevention: Assessment of Health and Care Impact

Introduction

The aim of this study is to assess the feasibility of applying the CanRisk risk prediction model, which includes the PRS, for risk stratification and breast cancer prevention in a real clinical setting at the Fondazione Policlinico Universitario Agostino Gemelli (FPG) in Italy, and to provide preliminary evidence on its clinical impact. Women visiting the Genetics Clinic for breast cancer screening will be enrolled in the study. They will undergo a blood sample withdrawal, which will be used to measure their polygenic risk score. The PRS will be included in the CanRisk evaluation to assess their cancer risk. After receiving their risk assessment, participants will be asked to complete a survey to provide feedback on their experience with the screening process, the blood sample collection, and their understanding and perception of the PRS and CanRisk evaluation. Healthcare providers involved in the study, including genetic counselors and physicians, will also complete a separate survey to share their insights and opinions on the feasibility, utility, and overall process of integrating PRS into breast cancer risk assessment. The primary outcome of the study is the feasibility and acceptability of incorporating PRS into the CANRISK evaluation, while secondary outcomes include participant satisfaction, understanding of the PRS and CanRisk evaluation, and healthcare provider feedback on the process. The study will be conducted per ethical guidelines, and informed consent will be obtained from all participants.

Data Collection for Algorithm Development

All enrolled women will undergo PRS assessment. Results in aggregate form will be provided to study participants upon request only. PRS will be performed on blood samples taken at the time of the visit. To extract the appropriate amount of DNA (100-200 ng) you will need at least 0.5 ml of blood. The selection of genotyped SNPs for inclusion in the PRS will be based on the model developed by Mavadatt et al., which included 313 SNPs, and subsequently validated in several other studies.

The demographic and clinical characteristics of the study participants will be summarized using descriptive statistics. Questionnaire data will be analyzed using descriptive statistics.

Technical Details of the Algorithm and Its Application

All the information regarding quality control and validation of the PRS algorithm could be found above (please see chapter “*Polygenic Risk Score algorithm, quality control and validation*”).

Testing will be performed with Gene Titan Thermofisher with the Axiom PMDA (Precision Medicine Diversity Array) kit (not validated for diagnosis). Through the use of Axiom™ Analysis Suite, Applied Biosystems™ Analysis Power Tools and the SNPolisher™ package, we will perform a quality control (QC) analysis, applying filtering criteria at both the SNP and sample levels. After the initial QC, the missing genotypes will be imputed using the HRC (Haplotype Reference Consortium) reference panel in the Helmholtz Munich Imputation Server, with the aim of improving the coverage of genetic variants and increasing the power of subsequent analyses. Following imputation, a post-imputation quality control will be applied to ensure the reliability of the imputed genotypes, excluding variants with low imputation quality scores, high genotyping uncertainty or deviations from the Hardy-Weinberg equilibrium. In addition, SNPs with lower low allele frequencies (MAFs) will be eliminated, as they are subject to higher error rates and potentially less informative for analysis.

For women included in the study, the following data will be collected: Age at enrollment; Age at menarche; Body mass index (BMI); Parity; Nursing; Breast density; Use of oral contraceptives; Menopausal status; Age at menopause (if applicable); Use of hormone replacement therapy (HRT) in menopause; Personal and family history of cancer (particularly breast, ovarian, endometrial, and gastrointestinal); Diagnostic test result, if performed; PRS Test Result.

In addition, women diagnosed with breast cancer will be required to provide the following additional information: Age at diagnosis; Surgery (yes/no); Type of surgery; Residual tumor (yes/no); Histology: histotype, grade, pathologic stage, chemotherapy response score (CRS) when applicable; Treatment: type of chemotherapy, number of cycles, response to treatment, maintenance therapy follow-up: patient status (cancer-specific death, death from other causes, alive), disease status (absent, stable disease, progressive disease), Disease-Free Survival (DFS), Overall Survival (OS).

The CANRISK model is an advanced tool used to calculate an individual's risk of developing breast and ovarian cancer. It incorporates a variety of variables, including:

- Personal Details: Sex, birth year, height, and weight (used to calculate BMI).
- Lifestyle Factors: Alcohol consumption.

- Reproductive History: Age at menarche, age at menopause, parity (number of children), and age at first live birth.
- Hormonal Factors: Use of oral contraceptives and menopause hormone therapy.
- Medical History: Age of diagnosis of breast, ovarian, prostate, or pancreatic cancers.
- Breast Screening: Mammographic density (BI-RADS classification).
- Genetic Testing: Results of genetic tests for high- and moderate-risk genes, including BRCA1, BRCA2, PALB2, CHEK2, ATM, RAD51C, RAD51D, BARD1, and BRIP1.
- Family History: Detailed family history of breast, ovarian, and other related cancers.
- Polygenic Risk Score (PRS)

The PRS is crucial in refining the risk assessment by providing a more personalized evaluation based on the individual's genetic makeup. By integrating the PRS, the CanRisk model can significantly alter the estimated risk, offering a more precise prediction and enabling better-informed decisions regarding cancer prevention and screening strategies

We will offer participation in the study to all eligible patients who will be referred to the involved department during the one-year study period, from 01/01/2025 to 31/12/2025, with an expected total of 700 subjects. We conducted the sample size calculation to ensure sufficient power to detect any change in diagnostic pathway timing before and after the introduction of PRS within the CanRisk model. A sample size of 510 subjects after modification provides 80% power to reject the null hypothesis of equal means. We performed this calculation with a significance level (alpha) of 0.05, using a two-sample bilateral t-test with equal variance. We used the PASS software to perform the sample size calculation.

For continuous variables, such as age and BMI, the data will be presented as mean values accompanied by their standard deviations (\pm SD) or as medians with their corresponding interquartile ranges (IQR), depending on the distribution of the data. Categorical variables, such as gender, family history of cancer, and genetic test results, will be expressed as absolute frequencies (counts) and percentages. A sensitivity analysis will be conducted using multiple imputation for missing data. Missing data patterns will be analyzed to assess potential biases. Analyses will be conducted using STATA version 17 or higher.

5.4. PRE-PDAC, Evaluation of Polygenic Risk scoreE for Pancreatic Ductal AdenoCarcinoma risk prediction: a case-control study

Introduction

The aim of the pilot is to evaluate the effectiveness of the polygenic risk score (PRS) in predicting pancreatic ductal adenocarcinoma (PDAC) risk, contributing to the development of more accurate predictive models that could improve clinical management and the prevention of PDAC.

The primary objective is to evaluate the association between the PRS, derived from the combination of known risk-associated SNPs, including ABO alleles, and the risk of developing PDAC. The primary endpoint will be the odds of developing PDAC across different PRS percentiles.

As a secondary objective, we will construct a multifactorial risk score by combining the weighted PRS PDAC with two well-established risk factors (smoking and diabetes). Additionally, within the PDAC patient cohort, we will evaluate the association between the PDAC PRS and clinical parameters, including pathological characteristics and cancer management strategies (such as surgery and medical treatments). Secondary endpoints include the odds of developing PDAC across multifactorial risk score percentiles as well as the distribution of clinical and pathological characteristics and cancer management approaches according to PRS percentiles.

Cases will be defined as patients with histologically confirmed PDAC, while controls will be individuals without a personal history of PDAC within the past ten years.

Data Collection for Algorithm Development

All enrolled patients will undergo PRS assessment. Results in aggregate form will be provided to study participants upon request only. PRS will be performed on blood samples taken at the time of the visit. To extract the appropriate amount of DNA (100-200 ng) you will need at least 0.5 ml of blood. The selection of genotyped SNPs for inclusion in the PRS will be based on the model developed by Mavadatt et al., which included 313 SNPs, and subsequently validated in several other studies.

The demographic and clinical characteristics of the study participants will be summarized using descriptive statistics. Questionnaire data will be analyzed using descriptive statistics.

Technical Details of the Algorithm and Its Application



All the information regarding quality control and validation of the PRS algorithm could be found above (please see chapter “*Polygenic Risk Score algorithm, quality control and validation*”).

In this case-control study two types of PRS will be calculated: an unweighted PRS and a weighted PRS. The unweighted PRS will be computed by summing the total number of risk alleles for each individual (assigning a value of 1 to each risk allele) and incorporating ABO blood group values, with 0 assigned for group OO, 1 for OA/OB, and 2 for group AB. The weighted PRS will be calculated as the sum of an individual's risk alleles, weighted by the odds ratios (ORs) derived from GWAS data on PDAC, with similar weighting applied for ABO blood groups (7).

Patients' characteristics will be described as absolute frequencies and percentages for categorical variables and as medians with ranges or means with standard deviations for continuous variables, as appropriate. Differences between groups (cases and controls) will be assessed using the chi-square test or Fisher's exact test for categorical variables, while numerical variables will be tested for normality using the Shapiro-Francia test, followed by Student's t-test or the Mann-Whitney U test, based on the normality assessment.

For both cases and controls, the following data will be collected: Sex; Age at enrolment; Smoking; Alcohol consumption; BMI (Body Mass Index); Personal and family history of cancer (particularly pancreas and breast cancer); Personal history of pancreatic diseases, diabetes, hypertension, periodontitis, ulcers (gastric or duodenal), Helicobacter pylori infection; Previous surgical interventions (such as gastrectomy and cholecystectomy); Pharmacological history (use of aspirin, NSAIDs, statins, PPIs, ACE inhibitors, ARBs); PRS. Only for cases the following additional information will be requested: Diagnosis: age at diagnosis, Ca19-9 and CEA levels, primary localization (head versus body-tail); Surgery: type of surgery; Histology: stage (defined according to NCCN as resectable, borderline resectable, locally advanced, or metastatic [21]); Treatment: chemotherapy, radiotherapy, EUS-guided ablative therapy.

Logistic regression will be employed to assess the association between individual SNPs and the outcome (PDAC: yes/no). The relationship between ABO blood groups derived from genotypes and PDAC risk will also be evaluated using logistic regression, with blood group O serving as the reference category.

The unweighted PRS and the weighted PRS will be assessed as continuous and categorical variables. The PRS will be categorized into the following percentiles based on its distribution in the control group: [0,5%), [5%,10%), [10%,20%), [20%,40%), [40%,60%), [60%,80%), [80%,90%), [90%,95%), and [95%,100%]. The lowest 5% ([0,5%]) will represent the lowest PRS group, while the highest 5% ([95%,100%]) will represent the

highest PRS group. The middle percentile range ([40%,60%]) will be used as the reference category.

Multifactorial risk scores will also be calculated by integrating the weighted PRS with tobacco smoking and diabetes variables. The calculated scores will be analyzed for their association with PDAC risk using logistic regression.

All analyses will adjust for the following variables: age, sex, Ca19-9 baseline levels, primary tumor localization and stage.

Finally, the frequency of the following variables will be calculated based on PRS percentiles: age, sex, stage, Ca19-9 baseline levels, primary localization, surgery, chemotherapy, radiotherapy, and EUS-guided ablative therapy. Statistical tests will then be performed to evaluate any potential associations.

All analyses will perform using *Stata software* (STATA/BE 17.0 for Windows, StataCorp LP, College Station, TX 77845, USA), with a statistical significance level set at $p < 0.05$.

References

1. Khera A V., Chaffin M, Aragam KG, Haas ME, Roselli C, Choi SH, et al. Genome-wide polygenic scores for common diseases identify individuals with risk equivalent to monogenic mutations. *Nat Genet* 2018 509 [Internet]. 2018 Aug 13 [cited 2025 Apr 15];50(9):1219–24. Available from: <https://www.nature.com/articles/s41588-018-0183-z>
2. Dareng EO, Tyrer JP, Barnes DR, Jones MR, Yang X, Aben KKH, et al. Polygenic risk modeling for prediction of epithelial ovarian cancer risk. *Eur J Hum Genet* [Internet]. 2022 Mar 1 [cited 2025 Apr 15];30(3):349–62. Available from: <https://pubmed.ncbi.nlm.nih.gov/35027648/>
3. Chatterjee N, Shi J, García-Closas M. Developing and evaluating polygenic risk prediction models for stratified disease prevention. *Nat Rev Genet* [Internet]. 2016 Jul 1 [cited 2025 Apr 15];17(7):392–406. Available from: <https://pubmed.ncbi.nlm.nih.gov/27140283/>
4. Dudbridge F, Wray NR. Power and Predictive Accuracy of Polygenic Risk Scores. *PLOS Genet* [Internet]. 2013 [cited 2025 Apr 15];9(3):e1003348. Available from: <https://journals.plos.org/plosgenetics/article?id=10.1371/journal.pgen.1003348>
5. Anderson CA, Pettersson FH, Clarke GM, Cardon LR, Morris AP, Zondervan KT. Data quality control in genetic case-control association studies. *Nat Protoc* [Internet]. 2010 Aug 26 [cited 2025 Apr 15];5(9):1564–73. Available from: <https://pubmed.ncbi.nlm.nih.gov/21085122/>
6. Das S, Forer L, Schönherr S, Sidore C, Locke AE, Kwong A, et al. Next-generation genotype imputation service and methods. *Nat Genet* [Internet]. 2016 Oct 1 [cited



2025 Apr 15];48(10):1284–7. Available from:
<https://pubmed.ncbi.nlm.nih.gov/27571263/>

7. Galeotti AA, Gentiluomo M, Rizzato C, Obazee O, Neoptolemos JP, Pasquali C, et al. Polygenic and multifactorial scores for pancreatic ductal adenocarcinoma risk prediction. *J Med Genet* [Internet]. 2021 Jun 1 [cited 2025 Apr 15];58(6):369–77. Available from: <https://pubmed.ncbi.nlm.nih.gov/32591343/>