

Publishable summary

The Work Package 3 of the DARE initiative, titled “Digitally-enabled Biomarker Discovery”, is part of the Spoke 3 “DIGITALLY-ENABLED SECONDARY AND TERTIARY PREVENTION” and includes 14 pilot studies focused on the identification and validation of specific biomarkers useful for secondary and tertiary prevention, together, in some cases, with the development of machine learning models able to support prediction of outcomes of clinical interest.

The 14 pilot studies are divided between 4 tasks, each targeting a specific clinical area and collectively addressing a broad range of predictive goals. Task 3.1 focuses on children and frail older adults, applying AI and machine learning to predict the risk of infections and acute adverse events in elderly frail frequent users of the Emergency Department, as well as to stratify vaccine response—contributing to improved vaccination strategies and early risk detection in vulnerable populations. Task 3.2 addresses oncology, aiming to predict clinical cancer phenotypes by integrating heterogeneous data types (e.g., somatic alterations, transcriptomic profiles, protein interaction networks) into interpretable models, with pilots dedicated to colorectal, lung, and myeloma cancers. Task 3.3 concerns cardiometabolic diseases, developing models to anticipate adverse events in chronic metabolic conditions such as type 2 diabetes and non-alcoholic fatty liver disease. Finally, Task 3.4 focuses on psychiatric and cognitive disorders, identifying pathological alterations such as brain age connectivity scores and markers of intellectual disability in genetic syndromes and subclinical psychiatric conditions, with the goal of predicting progression and enabling earlier, more personalised interventions.

The present deliverable deals with the preliminary evaluation of the tools developed so far. Assessment is performed by using machine learning assessment techniques. In doing so, particular attention is given to model generalization, i.e. the ability of a model to perform well on previously unseen data, by carefully tuning hyperparameters and evaluating performance on independent validation sets. Feature selection is also applied to identify the most informative variables, reducing model complexity and the risk of overfitting. To ensure the integrity of the evaluation process, strict protocols are followed to prevent data leakage. These practices contribute to building reliable and reproducible models.

The present document describes, on a task-by-task chapter basis, the current status of the models in terms of objectives, device usage, and tool development. For certain pilot studies that are already at an advanced stage of development, details on preliminary results and validation protocols (presented via a technical robustness and machine learning transparency form) are reported in the Appendix for the sake of clarity and readability.