Data Collection and ML Pipeline for Cardiovascular Risk Assessment

Ruggero G. Bettinardi

1. **Preamble:** Some known cardiovascular (CV) Risk Factors are already collected during the scan session (demographics, lifestyle information, heart rate, breathing rate, brachial and ankle blood pressure, significant blood markers like hemoglobin, hsCRP, total cholesterol, LDL, HbA1c). This is crucial to the following strategy.

2. **Perform Literature Review**: Conduct a comprehensive literature review to identify additional cardiovascular (CV) risk factors and relevant covariates that should be collected and tracked if possible, as well as to consolidate knowledge about how to interpret the collected known CV risk factors. Ensure the data collection design includes measurements for as many identified factors as possible.

3. **Unsupervised Clustering of Collected Known CV Risk Factors**: Perform unsupervised clustering analysis on the collected CV risk factors data prior to the cardio biosignals scan. This helps identify potential subgroups or clusters within the population based on their risk profiles.

4. **Unsupervised Clustering of Cardio Biosignal Data**: Apply unsupervised clustering techniques to the multimodal unlabelled biosignal data obtained from the scan. This step aims to discover inherent patterns or groups within the biosignal data that may correspond to different CV risk levels.

5. **Define "Soft" CV-Risk Classes**: Utilize the insights gained from the literature review and the clustering results to define preliminary or "soft" CV-risk labels. These labels can be binary (e.g., risk/no-risk) or multi-class (e.g., low-risk, medium-risk, high-risk) and serve as a basis for further analysis.

6. **Design a Supervised Algorithm**: Develop a supervised machine learning algorithm that can predict the "soft" CV-risk labels assigned to each subject using only the cardio biosignal data collected during the scan. This step assesses whether the cardio biosignal data alone can accurately classify individuals into different risk categories.

7. **Actionable Suggestions and Tracking**: If the supervised algorithm is successful, use its predictions to provide actionable health suggestions to subjects post-scan ("do more sport", "eat more/less XYZ", "see doctor"). Additionally, keep a record of the recommended actions for future evaluation and model validation purposes.

8. **Handling Unsuccessful Predictions**: If the supervised algorithm fails to accurately predict the soft CV-risk class, investigate alternative solutions such as refining the data collection process, improving the clustering methods, or incorporating additional data sources to enhance the model's performance, such as prompting feedback from clinicians.

9. **Longitudinal Data Collection**: Repeat all measurements, including biosignal and CV risk factors, on the same subjects annually.

10. **Use the aggregated historical data to build cardiovascular risk trajectories**. These trajectories will be used to predict future CV-risk labels and identify potential preventive actions to alter risk progression.

11. **Periodic Health Updates**: Ensure periodic updates on the general health status of subjects, especially if they miss subsequent scans. This includes tracking significant health events (e.g., mortality or other dramatic occurrences) to maintain the accuracy and relevance of the data, as well as the need to update the models.

12. **Integrate Medical Feedback**: As reliable medical feedback on the models becomes available, integrate this information to refine and improve the algorithms continuously.

13. **Semi-Supervised Learning with Annotations**: If clinicians can provide meaningful distinct CV risk classes and/or annotate at least a portion of the cardio biosignals, incorporate these into a semi-supervised learning framework. This approach leverages both labeled and unlabelled data to improve model training and performance.

14. **Bonus:** aggregate all longitudinal demographic, medical and other meta info with CV, Skin and other biosignals data to create an integrated “personalized health recommendation system”.