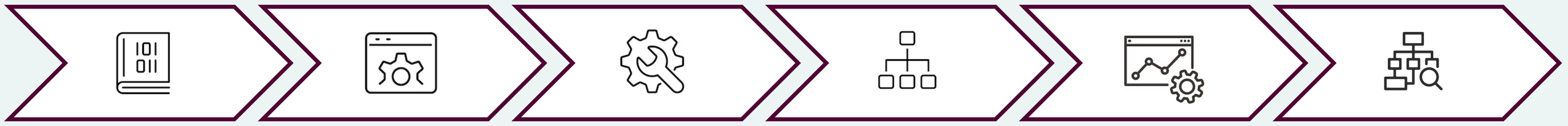
Using unsupervised machine learning methods of analysis for model variable selection in cancer risk prediction for PLHIV: a study protocol

Authors: Josefin Nilsson[1], Olof Elvstam[2,3], Isabela Killander Möller[1], Philippe Wagner[4], Johanna Brännström[1,6], Aylin Yilmaz[7,8], Fredrik Månsson[9], Magnus Boman[10,11], Christina Carlander[1,5,12]

Affiliations: [1]Unit of Infectious Diseases and Dermatology, Dept. of Medicine Huddinge, KI [2]Dept. of Translational Medicine, Lund University [3]Dept. of Infectious Diseases, Växjö Central Hospital [4]Center for Clinical Research, Västmanland, Uppsala University [5]Dept. of Infectious Diseases, Karolinska University Hospital [6]Dept. of Infectious Diseases, Södersjukhuset [7]Dept. of Infectious Diseases, Institute of Biomedicine, University of Gothenburg, Sahlgrenska Academy [8]Dept. of Infectious Diseases, Sahlgrenska University Hospital, Västra Götaland [9]Dept. of Clinical Sciences, Lund University, Infectious Diseases Research Unit [10]Dept. of Medicine Solna, Clinical Epidemiology, KI [11]MedTechLabs, Bioclinicum, Karolinska University Hospital [12]Dept. of Medical Epidemiology and Biostatistics, KI



Data acquired from the National HIV Registry Sweden (InfCareHIV), LISA, Population Registry, Migration Registry, Death Registry, Prescribed Drug Registry and the Patient Registry

Data pre-processed and cleaned

Optimal number of clusters calculated using the CH-index and silhouette score

K-means clustering and hierarchical clustering methods implementation

Dimensionality reduction by PCA, UMAP and t-SNE techniques implementation

Decision trees used to aid interpretation of the clusters and relationships to determine the variable importance

Conclusion

Using an exploratory methodology integrating multiple analytical techniques, the results can be used optimise variable utilisation in cancer prediction models for people living with HIV (PLHIV).

This approach can enhance the understanding of risk factors for specific cancer types in PLHIV to improving prediction of cancer and accuracy.

Background

Risk of developing certain cancers is higher and may appear at a younger age among PLHIV

Yet there is a lack of guidelines for cancer screening recommendations specifically for PLHIV

Unsupervised machine can enhance model development and provide deeper insight into variable characteristics and relationships, recognising patterns which may have been overlooked by conventional statistical analysis methods

Aims

The aim of the study is to analyse and compare variable associations using a variety of unsupervised machine learning elevate understanding and and improve the selection of variables in the model to enhance cancer prediction.

Future plans

The goal is to use the information gained about the variables through this analysis to build risk prediction models for clinical use using supervised machine learning methods

Josefin Nilsson, MPH, PhD Student

Karolinska Institutet
Departement of Medicine Huddinge
josefin.nilsson@ki.se

Karolinska Institutet