A Random Forest Algorithm for Identifying Risk Factors for Multimorbidity in the UK Biobank Cohort

Patel Linia[1], Mignozzi Silvia[1], Pizzato Margherita[1], La Vecchia Carlo[1], Alicandro Gianfranco[2, 3]

(1) Department of Clinical Sciences and Community Health, Università degli Studi di Milano, Milano, Italy
(2) Department of Pathophysiology and Transplantation, Università degli Studi di Milano, Milano, Italy
(3) Mother and Child Department, Cystic Fibrosis Centre, Fondazione IRCCS Ca' Granda Ospedale Maggiore Policlinico, Milano, Italy

CORRESPONDING AUTHOR: *Alicandro Gianfranco, gianfranco.alicandro@unimi.it*

INTRODUCTION

High-income countries are undergoing significant demographic shifts, characterized by population decline and progressive aging. These transformations are associated with an increase in the prevalence of chronic diseases, which often coexist, worsening individuals' quality of life and increasing healthcare costs. Identifying the factors that contribute to the onset of multimorbidity is particularly complex, as these factors often interact with each other and cause multiple effects across different diseases.

OBJECTIVES

This study aimed to identify the main risk factors for multimorbidity within a large UK cohort using a fully nonparametric ensemble method. This approach makes no assumptions about the underlying relationships between variables and allow managing high-dimensional data while preventing overfitting.

METHODS

We analyzed data from the UK Biobank cohort, which includes detailed information on socioeconomic status, lifestyle, anthropometric measures, and environmental exposures collected at recruitment, along with disease occurrence obtained through linkage with hospital admissions (primary and secondary diagnoses), death records, and cancer registries. Multimorbidity was defined as the presence of at least two chronic conditions from a list developed through an international consensus using a modified Delphi method [1]. To assess the role of 18 candidate variables in predicting the onset of multimorbidity over a five-year follow-up, we applied a random forest algorithm adapted for survival analysis within a competing risk framework [2], considering two competing events: the development of multimorbidity and death prior to its onset. The candidate variables included: white British/Irish ethnicity (Yes/No), qualification level, average total household income before tax (adjusted for household size and categorized into quintiles), area-level index of multiple deprivation (deciles), body mass index (kg/m2), waist circumference (cm), pack-years of smoking, alcohol drinking (g/day), healthy diet score (ranging from 0 to 5, based on the intake of fruit, vegetables, fish, whole grains, processed and red meat), walking (at least 10 min, number of times a week), moderate physical activity (at least 10 min, number of times a week), vigorous physical activity (at least 10 min, number of times a week), particulate matter air pollution 2.5 (PM2.5) (µg/m3), PM2.5-10 (µg/m3), PM10 (µg/m3), NO2 (µg/m3), average exposure to evening (7:00 pm – 11:00 pm) or night noise (11:00 pm – 7:00 am) (dB). Results were summarised using out-of-bag partial dependence plots and variable importance (VIMP) metrics.

RESULTS

Of the 422,344 individuals included in the cohort, aged between 39 and 73 years, we selected 137,565 participants who were free from the conditions included in the definition of multimorbidity at the time of recruitment and for whom risk factor information was available. During the five-year follow-up, 4384 individuals developed multimorbidity (2740 males, 1644 females). The five-year cumulative incidence was 3.9% in males and 2.6% in females. Among individuals who developed multimorbidity during follow-up, the main conditions observed were cancer (52.4% of males and 52.1% of females), arrhythmias (44.7% of males and 28.5% of females) and coronary artery disease (42.1% of males and 24.8% of females). Based on VIMP metrics, the strongest predictors in men were smoking, waist circumference, and sleep duration; in women

DOI: 10.54103/2282-0930/29299

alcohol, smoking, and waist circumference. Five-year cumulative incidence was higher for heavy smokers (sex-specific 95th percentile of pack-years) (males: 6.3%, females: 4.0%) compared to non-smokers (males: 3.5%, females: 2.4%); for individuals with elevated waist circumference (sex-specific 95th percentile) (males: 6.1%, females: 5.2%) versus those with median values (males: 3.9%, females 2.6%); for heavy alcohol drinkers (sex-specific 95th percentile) (males: 4.6%, females: 4.0%) versus median intake (males: 3.8%, females: 2.4%); for those sleeping 4 hours/day (males: 6.3%, females: 4.2%) or 10 hours/day (males: 6.5%, females: 4.5%) versus 7 hours/day (males: 3.7%, females: 2.5%). Diet, physical activity, and air pollution had smaller impacts.

CONCLUSIONS

Preventive interventions targeting smoking, abdominal obesity, and heavy alcohol consumption among middle-aged adults in the UK and likely in other high-income countries, may substantially reduce the incidence of multimorbidity. Such interventions could improve the health trajectory and burden of disease of future older populations. In addition, promoting adequate sleep duration appears to be beneficial and should be integrated into public health recommendations.

REFERENCES:

1. Ho ISS, Azcoaga-LorenzoA, Akbari A, et al. Measuring multimorbidity in research: Delphi consensus study. BMJ Med 2022 Jul 27;1(1):e000247. doi: 10.1136/bmjmed-2022-000247.
2. Ishwaran H, Gerds TA, Kogalur UB et al. (2014) Random survival forests for competing risks. Biostatistics 15, 757-773.

DOI: 10.54103/2282-0930/29299

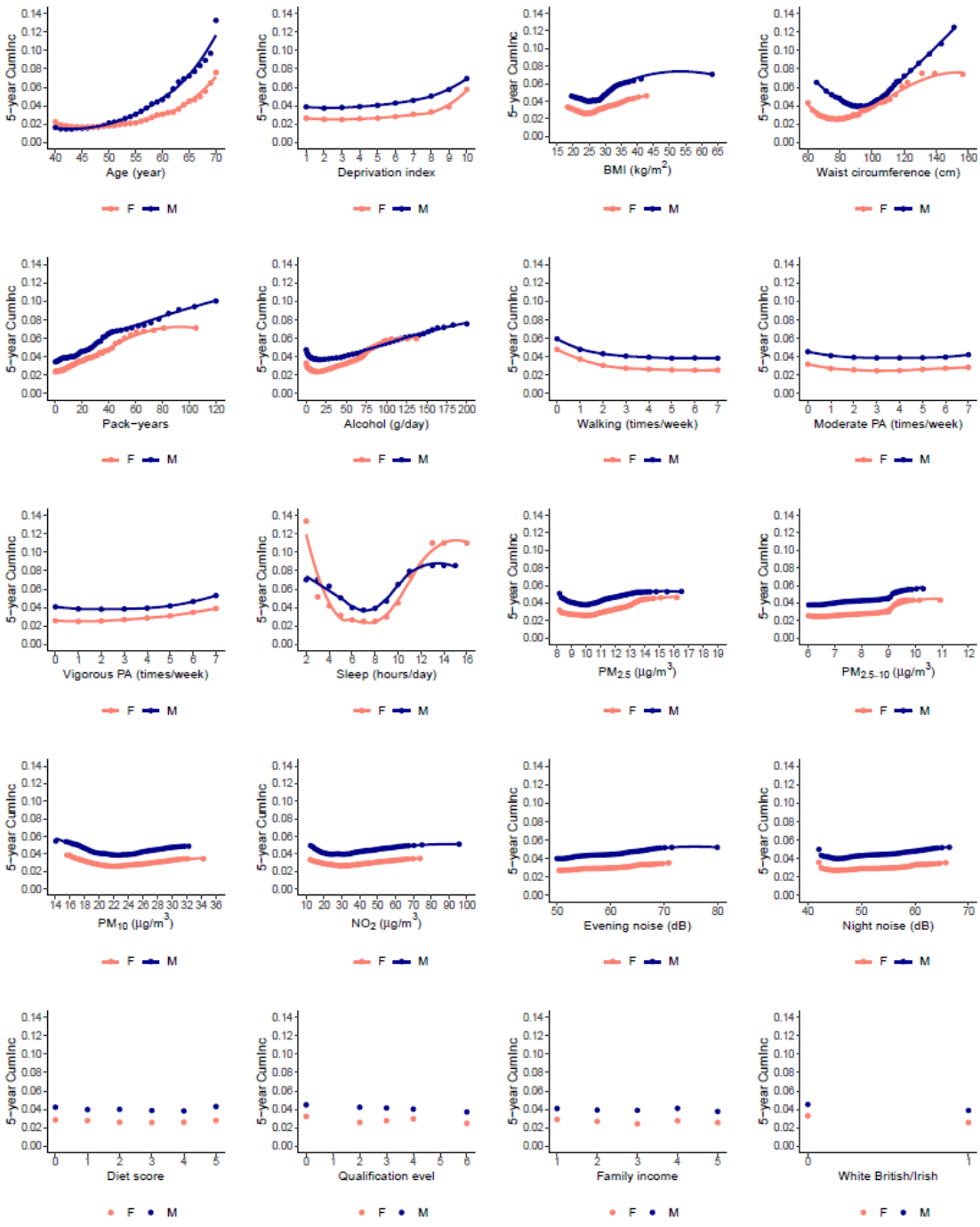

Figure 1. Partial dependence plots displaying the predicted five-year cumulative incidence of multimorbidity across different levels of socioeconomic, lifestyle and environmental exposures, stratified by sex. Family income before tax was adjusted for household size and categorized into quintiles. Diet score, ranging from 0 to 5 (with 5 indicating the healthiest diet), was based on intake of fruit, vegetables, fish, whole grains, and consumption of processed and red meat. Qualification levels are based on the UK Regulated Qualifications Framework (RQF), which classifies education attainment from Entry Level through Level 8 [Level 0: No qualification; Level 1: General Certificate of Secondary Education (GCSE) or equivalent functional skills; Level 2: Higher attainment (e.g. GCSE grades A–C); Level 3: A-levels and access diplomas; Levels 4–5: Sub-degree higher education qualifications (e.g. HNC, foundation degrees); Level 6: Undergraduate degrees; Level 7: Postgraduate qualifications (e.g. Master's, PGCE); Level 8: Doctoral-level education (e.g. PhD)].

DOI: 10.54103/2282-0930/29299