International Journal of Population Data Science

Journal Website: www.ijpds.org

Feasibility research into using administrative data sources to produce multimorbidity scores for predicting general health statistics for England

Tolu Adedire[1], Wajiha Munir[1], Charlotte Standeven[1], and Matthew Minifie[1]

[1]Office of the National Statistician, Manchester, United Kingdom

The current primary measure of general health in England relies on census and survey data, which are limited in timeliness and granularity. Developing an administrative data-based indicator offers the potential for more frequent, detailed insights to support research and policy. This paper sets out research to develop and evaluate a multimorbidity score derived from linked administrative health datasets, combining National Health Service (NHS) hospital records, NHS General Practice data, and the Office for National Statistics' death registrations, which are then joined to census information for residents of England. Moving beyond simple counts of conditions, or prescriptions, the multimorbidity score provides a structured measure of the burden and complexity of chronic illnesses at the individual level. We explore using this score as a key predictor of health-related outcomes, including self-reported general health status. Machine learning models were trained to predict the general health status responses from the 2021 Census. Given the subjective nature of census responses, we considered exposures on clinical indicators, healthcare utilisation metrics, demographic factors, social factors, and lifestyle factors. To ensure robustness, we experimented with multiple modelling approaches, including tree-based algorithms and regression-based methods, comparing their predictive performance across evaluation metrics. Model performance was assessed on an independent subsample using multiple evaluation metrics, and predicted probabilities were aggregated to produce breakdowns by key characteristics, which were compared against observed values. Finally, the paper considers the feasibility of generating a time series from 2015 onwards and discusses potential applications and limitations of these estimates for producing broader health-related statistics.

https://doi.org/10.23889/ijpds.v11i5.3495