Shared predictors of social participation in psychosis and the general population: The central relevance of emotional and relational wellbeing

Shared Predictors of Social Participation in Psychosis and the General Population:

The Central Relevance of Emotional and Relational Wellbeing

By: Dr. Fergus Quilligan

Clinical Neuroimaging Laboratory & Centre for Neuroimaging, Cognition and Genomics (NICOG),

Galway Neuroscience Centre,

School of Medicine,

University of Galway,

Ireland.

Submitted for M.Sc.

September 2025

Table of Contents

Acknowledgements

I would like to begin by sincerely thanking my supervisor, Prof Dara Cannon, and co-supervisor Prof. Derk Morris, for the opportunity they gave me by taking on an applicant from left field to participate in this work, and her invaluable guidance and support throughout this project. I am also grateful to my committee/examiners, for their valuable time, insights, reading and feedback of this work.

I would like to thank my lab-mates and colleagues on the project, as well as the postdocs, in particular Aodán, Mia, Evie and Emma, who gave someone with limited domain expertise a lot of time and support, both scientific and emotional, during this work.

Most importantly, I would like to thank my family, most especially my wife Emma, whose love and support made all of this feasible – from supporting the concept of doing this work, to giving me the space and time to deliver on it during procrastination-induced bursts of productivity, and to my friends for taking me away and providing the right distractions at the right time – you know who you are!

Funding: This research was funded by Taighde Eireann (Research Ireland) proposal number 21/FFP-P/10165. This research has been conducted using the UK Biobank Resource under Application Number 98153
Ethics approval: UK Biobank received ethical approval from the Northwest Multi-center Research Ethics Committee.
Consent to participate: All participants provided consent to the Uk Biobank at recruitment.
Availability of data and material: The dataset used is not publicly available as UK Biobank require a data access application however, information on how to access the data can be found here https://www.ukbiobank.ac.uk/use-our-data/apply-for-access/.
Code availability: Available upon request.

I would like to thank the participants of the UK Biobank whose volunteering of time and information made this research possible.

Abstract

Introduction:

Social participation (SP) lowers mortality rate and risk for depression, dementia (Wanchai & Phrompayak, 2019) and functional disability (Ashida et al., 2016) as well as being a factor influencing recovery from severe psychiatric illnesses such as psychosis (Harrison et al., 2001). This is due to complex combinations of biological, psychological and behavioural mechanisms. Demographic factors such as socioeconomic status have been shown to influence social participation (Kung et al., 2022) as well as more specifically focused studies implicate life factors such as urbanism, community connection, life satisfaction and anxiety. However, the relative strength of association of these factors on social participation remains unclear. Using a large-scale prospective population cohort, I aim to identify significant correlates of social participation and, furthermore, to examine whether these associations differ in psychosis.

Methods:

Social participation was derived from the frequency of leisure and friend/family activities (range 0-10) in the UK Biobank. The relative importance of a broad range of 514 demographic, psychosocial, health and lifestyle variables on social participation was established using XGBoost on a data sample, with predictive regression models developed to quantify the associations completed on an independent sample. The moderating effect of psychosis on these relationships was then analysed.

Results:

Regression analysis of the dataset (n=492,875, age mean years ± SD 56.5±8.1) revealed that, when combined with age and sex, satisfaction metrics including friendship satisfaction (R^2=0.067), perceived life meaning (R^2=0.051), ability to confide (R^2=0.047), happiness (R^2=0.046), and family relationship satisfaction (R^2=0.046) were among the strongest correlates of social participation, along with employment status (R^2=0.043). An eight-parameter model was generated to predict 11.5% of the variance in social participation (3.8% explained by sex/age). Factor analysis identified a single latent emotional and relational wellbeing factor that accounted for 8.35% of the variance in social participation (3.7% explained by sex/age). These relationships did not differ significantly between individuals with and without psychosis.

Conclusions:

Perceived emotional and relational wellbeing, along with employment status and access to transport are statistical predictors of the extent of social participation, both in a healthy population and for those with psychosis. These findings highlight factors that are related to better social integration for individuals in the general population and those with psychiatric disorders. These links can therefore help inform policy to provide more effective interventions, thereby reducing associated health and economic burdens.

Chapter 1: Introduction

Social Participation

Social Participation and Health

Social participation (SP) is defined as involvement in activities or interactions with others in community or interpersonal contexts. SP encompasses a broad range of behaviours, from informal socialising to structured volunteering and civic engagement. A growing body of research highlights the protective role of SP against various negative outcomes, positioning it as a social determinant of health (Holt-Lunstad et al., 2015; Mutz et al., 2021). Demographic change, rapid urbanisation, and growing social fragmentation have been identified as major public health challenges (Cacciatore et al., 2025) and are associated with elevated rates of psychosis and poorer mental health outcomes (Ku et al., 2021; Vassos et al., 2012). Identifying the factors that facilitate or hinder social participation has therefore become an important priority for public health and psychiatric research.

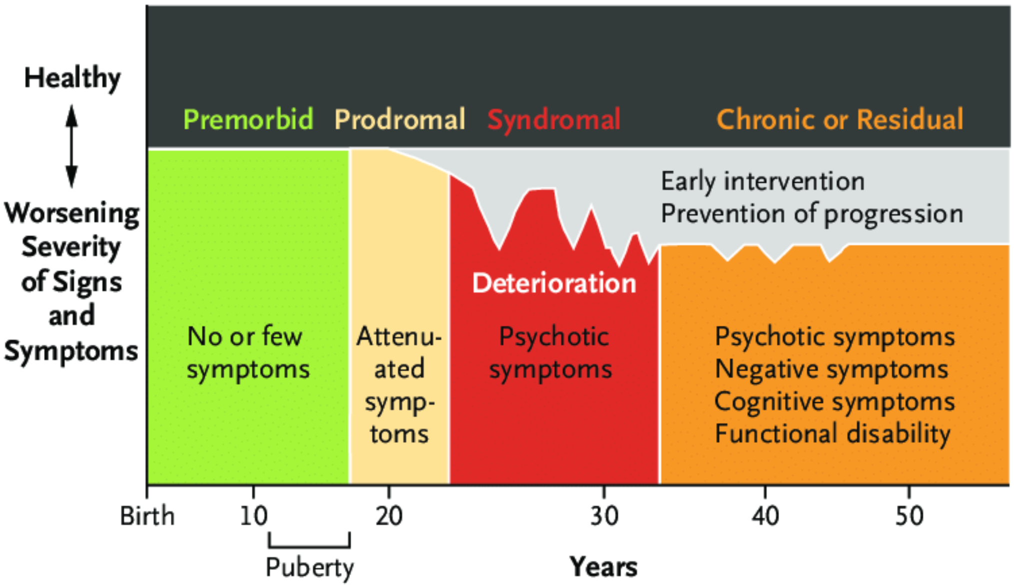
SP is consistently associated with adult mental health and wellbeing. Both informal relationships and structured community involvement reduce risks of depression, anxiety, and cognitive decline (Fratiglioni et al., 2004; Shen et al., 2022). Emotionally supportive relationships buffer stress and reduce vulnerability to depression and anxiety (Cohen & Wills, 1985). Longitudinal evidence indicates that SP is associated with increased risks for developing common psychiatric disorders, such as depression and often results in a poorer trajectory for those with these disorders (Santini et al., 2015; Holt-Lunstad et al., 2010).

Addressing social isolation and promoting SP is both a public health and a social equity imperative. The health consequences of social isolation, living alone and subjective loneliness involve an elevated mortality risk of 26-32% based on a meta-analysis comprising over 3 million individuals from 70 studies (Holt-Lunstad et al., 2015). These risks are comparable in magnitude to traditional risk factors such as obesity and light smoking (Flegal et al., 2013; Jacobs et al., 1999). Individuals with higher levels of social functioning have been found to have reduced incidence of functional disability, dementia, and all-cause mortality (Fratiglioni et al., 2004; Ashida et al., 2016; Foster et al., 2023). Better social functioning is also associated with improved cardiovascular function and immune response (Holt-Lunstad et al., 2010), while social isolation has been identified as an independent risk factor for conditions such as Parkinson’s disease (Geng et al., 2024). Policy frameworks from organisations such as the World Health Organisation (WHO) on aging (*Decade of Healthy Ageing*, 2021) emphasise participation as both a human right and a critical determinant of health trajectories. Accordingly, fostering SP should be a key target for health promotion and policy intervention, particularly as populations age and societal structures become more fragmented.

Social Participation and Psychosis

The issue of social participation is key in psychosis. People living with psychotic disorders often experience marked reductions in social engagement, with consequences for recovery, quality of life, and social inclusion. Many individuals treated for psychosis see partial recovery, but this is associated with functional impairments, stigma, and disruptions to social networks that contribute to persistent limitations in education, employment and social activities (Figure 1). While clinical factors (e.g., symptom severity, cognition) explain some variance, broader psychosocial and environmental determinants are increasingly recognised. Clarifying the shared and distinct predictors of participation in psychosis and in the general population is therefore crucial to inform targeted interventions.

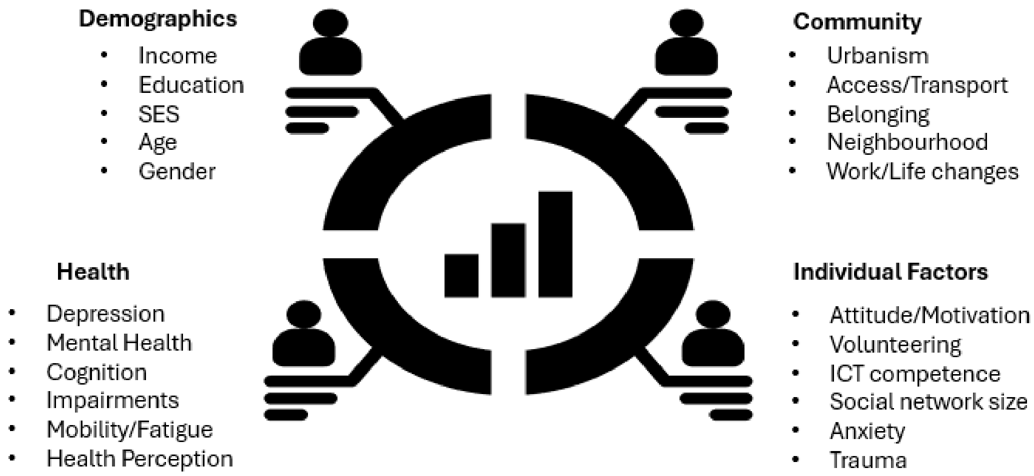
Figure 1: Sample Psychosis progression through premorbid up to chronic stages (Lieberman & First, 2018)



Determinants of Social Participation

Despite increasing recognition of its importance, the determinants of SP remain insufficiently understood. Demographic characteristics such as income, education, and socioeconomic status (Ashida et al., 2016; Kung et al., 2022) have been shown to shape opportunities for engagement, while community-level contexts including urbanicity, access to transport, and a sense of belonging (Levasseur et al., 2015; B. G. Townsend et al., 2021; Wanchai & Phrompayak, 2019) also play a role. Health-related variables (Repke & Ipsen, 2020) ranging from perceived health and mobility limitations to experiences of mental health difficulties, can either enable or restrict participation. In addition, individual-level influences such as personal attitudes, prior experiences of trauma, and anxiety are likely to affect both willingness and capacity to engage (Allen et al., 2023). Collectively this illustrates the multifactorial nature of SP. However, while each domain has been implicated (Figure 2), their relative importance and potential interactions remain unclear, underscoring the need for data-driven approaches to disentangle these contributions.

Figure 2. Demographic, Psychosocial, Health and Lifestyle Domains Associated with Social Participation



SP is especially critical for individuals with severe psychiatric disorders, especially those that demonstrate features such as psychosis. While research increasingly acknowledges the role of social integration in recovery (Linde et al., 2023), few studies have systematically examined how social factors influence outcomes in psychiatric populations (Karadzhov, 2023; Stickley & Wright, 2011). Given the disruptions to social functioning commonly observed in psychosis, it is important to assess whether the predictors of SP differ between those with and without psychosis (Bae et al., 2025). Crucially, these determinants do not act independently. Instead, they interact in nonlinear ways: for example, the effect of health limitations may be moderated by social support or neighbourhood resources; living in a rural location may be moderated by availability of public transport or the ability to drive. This complexity makes it challenging to disentangle the relative contribution of distinct factors using traditional linear models.

Aims and Objectives

Despite extensive work on the health benefits of social participation, several gaps remain in understanding the variables that are associated with social participation. In particular, research to date has tended to examine predictors from individual domains in isolation, such as sociodemographic, psychosocial, health, or lifestyle factors. Far fewer studies have integrated these domains simultaneously to explore how multiple dimensions of individual and contextual factors jointly shape social participation. Addressing this gap requires analytic approaches capable of handling complex, potentially non-linear relationships and interactions among diverse predictors.

Supporting this approach, the development of advanced analytical techniques such as machine learning that can break down the interactions and understand the non-linear nature of some of these interactions is a key capability that brings a superior level of methodological rigour to the present analysis. The combination of leveraging these novel approaches with the advent of large-scale data available through population-based databases such as the UK BioBank opens up huge opportunity to generate comprehensive insights in a way that has not been possible previously. Moreover, because the UK Biobank includes rich medical and psychiatric data, it offers a unique opportunity to investigate whether the factors associated with social participation differ for individuals with psychosis compared to the general population.

The central questions addressed in this thesis are therefore:

- Taking a data driven approach, which variables, across multiple domains examined in combination, are most strongly associated with social participation in the general population?
- Can machine learning methods such as XGBoost identify key predictors of social participation that might be missed by traditional approaches?
- How robust and interpretable are these predictors when examined within conventional regression frameworks?
- Do similar patterns of association emerge among individuals with psychosis, or are the predictors of social participation distinct in this group?

By addressing these questions, this work aims to advance understanding of the complex, multidomain influences on social participation and to explore their relevance in both population and clinical contexts.

Chapter 2: Methodological Framework

Introduction

The advent of large-scale prospective datasets that include an extremely broad variety of life factors, such as the UK Biobank, provide new opportunities to investigate multifactorial associations with SP. In parallel, the development and availability of machine learning (ML) approaches enable analysis of complex interactions in large datasets, providing an unbiased way to identify key variables. These approaches offer flexible alternatives that can capture nonlinearities and higher-order interactions without specification. Ensemble methods such as random forests and gradient boosting have demonstrated superior predictive performance across diverse health applications. Importantly, ML models can accommodate high-dimensional data, making them suitable for biobank-scale resources where hundreds of candidate predictors are available. The convergence of these capabilities provides the foundation for this novel investigation into the correlates of SP.

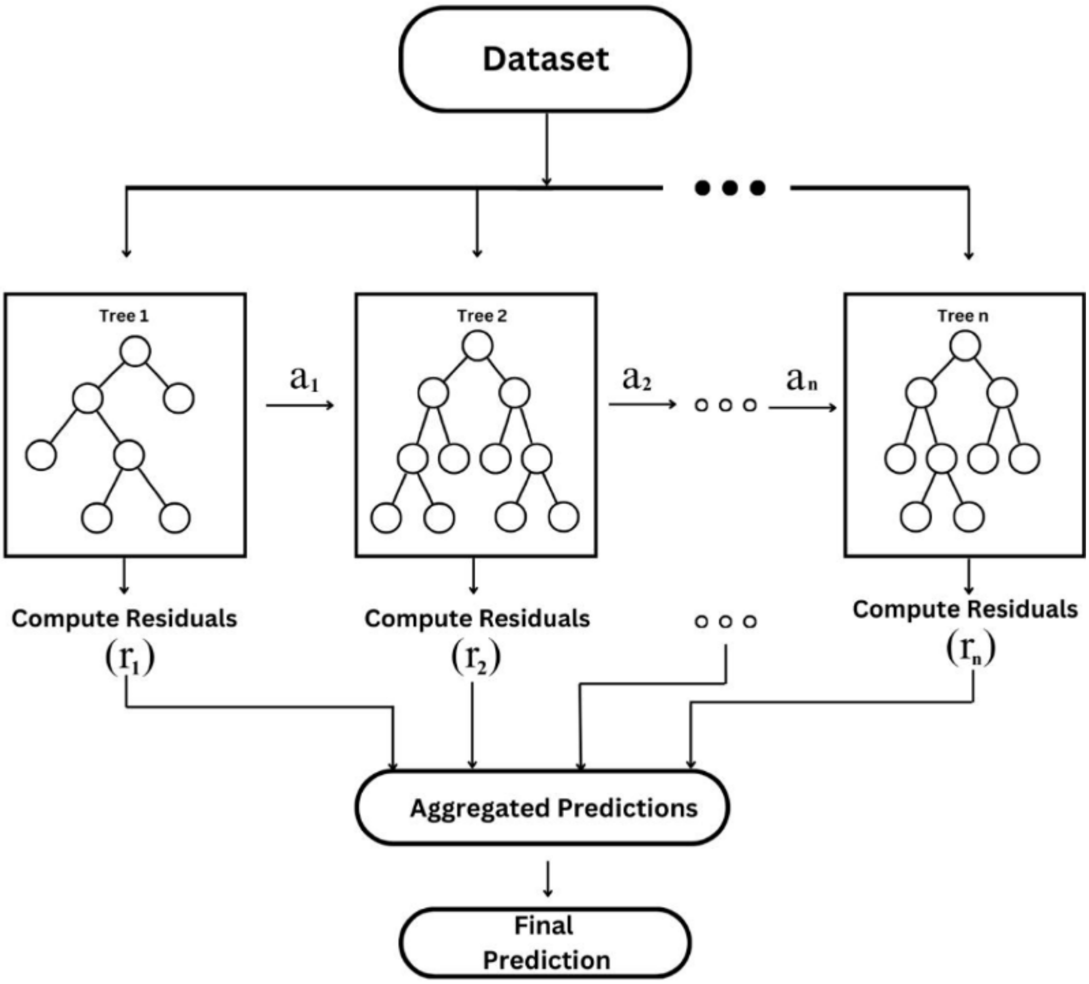
XGBoost

What is XGBoost?

Extreme Gradient Boosting (XGBoost) is an advanced implementation of the gradient boosting framework for supervised machine learning (Chen & Guestrin, 2016). Gradient boosting sequentially builds an ensemble of shallow decision tress referred to as weak learners, where each new tree is trained to predict the residual errors (gradients) of the combined model so far (Friedman, 2001). The ensemble thus "boosts" performance by progressively reducing the residuals of prior iterations, creating a strong predictive model from many weak learners (Figure 3).

Unlike bagging-based algorithms such as Random Forests, where trees are trained independently on resampled data, boosting algorithms build trees additively so each new tree explicitly corrects the errors of the previous ensemble. This sequential correction mechanism allows boosting models to achieve higher accuracy, particularly when complex non-linear relationships exist between predictors and outcomes.

Figure 3. Schematic diagram of XGBoost architecture. The variables a_n and r_n are the regularisation and residual parameters for the nth tree (Fry-Nartey et al., 2025)



Why use XGBoost?

XGBoost introduces several innovations that distinguish it from both earlier gradient boosting algorithms and other ensemble techniques:

1. Second-order gradient optimisation.
   Standard gradient boosting uses only the first derivative (the gradient) of the loss function to guide updates. XGBoost also incorporates the second derivative (the Hessian), which provides curvature information about the loss landscape. This second-order optimisation enables faster convergence, more stable learning, and improved performance, particularly on large, high-dimensional datasets, compared with classical Gradient Boosting Machines (GBMs), which rely solely on first-order updates (Chen & Guestrin, 2016)

2. Regularisation to prevent overfitting.
   XGBoost adds explicit regularisation terms (L1 and L2 penalties) to the objective function. Regularisation penalises overly complex models by

adding a cost for large or numerous parameters, thereby improving generalisability to unseen data.

- L1 regularisation (lasso) encourages sparsity by shrinking less informative weights to zero.
- L2 regularisation (ridge) discourages large weights but retains all features.
  Traditional gradient boosting methods, and most tree-based algorithms such as Random Forests, do not include these penalties by default. This built-in regularisation helps XGBoost achieve strong predictive performance without severe overfitting, even when hundreds of predictors are used.

3. Intrinsic handling of missing data.
   XGBoost can automatically manage missing values during tree construction by learning the optimal default direction (left or right branch) when a feature is missing. This design allows it to use incomplete observations directly, avoiding the need for prior imputation or deletion. In contrast, most classical algorithms—including linear regression, support vector machines, and neural networks—require explicit imputation or removal of missing data before training. The only assumption XGBoost makes about missingness is that the missingness is the same in the test and training sample.

Collectively, these innovations make XGBoost a robust, regularised, and efficient approach that preserves the interpretability of tree-based models while achieving accuracy comparable to or exceeding that of more complex methods such as deep neural networks. For these reasons, XGBoost is increasingly applied in health research, including disease risk prediction, feature selection, and integrative modelling of multidomain datasets.

The UK BioBank

The UK Biobank (UKB) is a unique resource that combines biomedical, lifestyle, environmental, and psychosocial data on over 500,000 middle-aged individuals. Its scale and breadth make it well-suited to the study of social participation and its determinants. The present study leverages this dataset to determine the strongest factors associated with SP across a broad adult population, including those with psychosis. However, its high dimensionality and substantial missingness pose analytic challenges. Machine learning approaches such as XGBoost are therefore well-matched to the data structure, offering an opportunity to model complex predictors while maintaining methodological rigour.

Methodological Considerations

One concern in high-dimensional modelling is the risk of "information leakage" — when the same dataset is used both for selecting features and for building final models, leading to inflated performance estimates. To address this, rigorous data-splitting strategies can be used, with independent subsamples allocated for feature selection and model fitting. While approaches vary, allocating a smaller portion of the sample for feature discovery (e.g., 20%) and the majority for model validation (e.g., 80%) helps reduce overfitting while maximising power in final models.

Drawing on over 500 demographic, psychosocial, health, and lifestyle variables, machine learning (XGBoost) was applied to identify the strongest contributors to SP on 20% of the data. Cross validated regression models were built using these variables on an independent data sample, in order to quantify their explanatory power, and conduct moderation analyses to compare to a psychosis cohort. This dual approach enables both variable selection and theoretical testing, yielding an objective, data-driven model of SP determinants across general and clinical populations.

Study Design

The aim of this study is to examine the effects of psychosocial, lifestyle, and demographic associations with social participation (SP), and to test whether these associations remain for individuals with psychosis, using data from the UK Biobank, a large, prospective population-based cohort study designed to investigate the genetic, environmental, and lifestyle determinants of a wide range of health outcomes.

Data were analysed to identify the most important demographic, psychosocial, health, and lifestyle variables, which were then used in regression models and the creation of a latent variable The resulting models and factors were subjected to a moderation analysis to ascertain if the relationships identified differ significantly in a psychosis cohort (Figure 4)

Figure 4. Summary of Methodological Approach and Analysis Pipeline Toward the Optimal Prediction of Variance in Social Participation

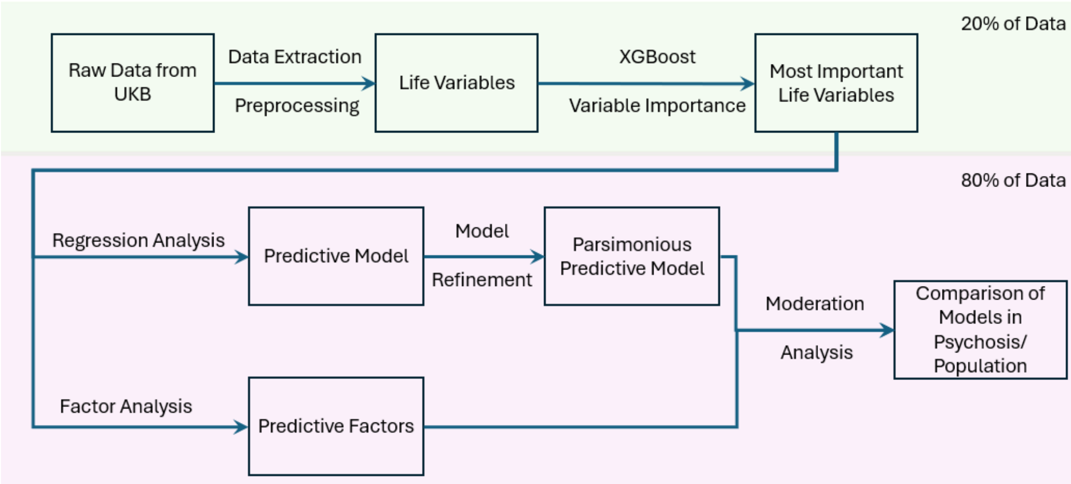


Figure 4 Legend: Summary of the analysis pipeline. The broad sample of features was dimensionally reduced using XGBoost, after which regression analysis and factor analysis were completed to deliver predictive models and latent factors. These models and factors were then compared for the general population and those with psychosis.

Participants

Between 2006 and 2010, 502,505 participants aged 40–69 were recruited across the United Kingdom. Participants provided informed consent for baseline assessments, including sociodemographic, psychological, and lifestyle measures, alongside cognitive and physical health evaluations. The UK Biobank received ethical approval from the North West Multi-centre Research Ethics Committee (REC reference: 16/NW/0274). Access to UKB data was granted under the application ID 98153.

Cohort Definitions

Psychosis Cohort Definition

The psychosis cohort was defined as participants who had an ICD-10 diagnosis (p41270) of any of the following conditions: F20-29 Schizophrenia, schizotypal and delusional disorders; F30.2 Mania with psychotic symptoms; F31.2 Bipolar affective disorder, current episode manic with psychotic symptoms; F31.5 Bipolar affective disorder, current episode severe depression with psychotic symptoms; F32.3 Severe depressive episode with psychotic symptoms; F33.3 Recurrent depressive disorder, current episode severe with psychotic symptoms. Additionally, participants reporting a professional diagnosis of schizophrenia or any other psychotic disorder (p29000) were included. This approach was taken to maximise the data cohort of those with psychotic disorders as the amount of individuals with a specific diagnosis of schizophrenia was small.

Social Participation Quantification

Social participation (SP) was examined as a composite of two domains, frequency of visits with friends and family (p1031) and leisure/social activities (p6160), which are objective, structural measures often used to measure SP (Foster et al., 2023; O'Connell et al., 2024). Frequency of visits with friends and family ranges from 0-5 based on an index of frequency. Leisure and social activities is encoded so participants receive a point for every activity they participate in including sports club/gym, pub/social club, religious group, adult education or other group activity, resulting in a score from 0-5. These data do not load onto a single latent variable in a manner that resulted in an acceptable fit, as they are evaluating distinct aspects of SP. However, these measures do align in the broader theoretical framing of SP, which led to the decision of using composite scoring. The frequency of visits was therefore summed with the quantity of leisure and social activities contributed to an overall SP index ranging from 0-10.

Data Selection & Preprocessing

The dataset was restricted to individuals with complete data on both SP components yielding a final sample of n=492,875 participants. The UKB Showcase tool was used to select the psychosocial, lifestyle, health and demographic data (Appendix 1Table S1.1). Variables with more than 70% missing values were excluded. A subset of these variables, such as ICD-10 diagnoses, was only available in a list format and was decoded into unique variables. Where unlimited options exist within a list, such as ICD-10 diagnoses, only patterns with more than 10,000 occurrences were decoded (Appendix 1 Table S1.2). Seven variables (Appendix 1 Table S1.3) required decoding to resolve special character values (e.g. -10 for <1 hour TV/day), and education data was transformed into years of schooling (Okbay et al., 2022).

Although XGBoost is robust to collinearity (Hastie et al., 2001a; Rojas et al., 2024), a correlation cutoff of 0.85 was nonetheless applied to reduce redundancy. Predictors with zero variance were also removed. To avoid information leakage during feature selection, a random 20% subsample was reserved for feature extraction; feature importance was estimated on this subsample, and the top predictors were modelled on the remaining 80%. This holdout-based selection strategy follows established procedures that use data splitting to screen and validate predictors for black-box learners (Tansey et al., 2022)

Identification of Most Important Variables

After preprocessing, many variables still contained substantial missingness. To address heterogeneous missingness, the dataset was partitioned into five subsets ('bags') based on the extent of missing data. XGBoost models were run on each subset, with 5-fold cross validation and 50 bootstrap iterations, using the *tidymodels* (Kuhn & Wickham, 2020) and *parsnip* (Kuhn & Vaughan, 2025) packages, with model parameters tuned using Bayesian optimisation. The summed importance values for the variables were extracted. Feature selection was performed per bootstrapped sample and the top 15 features per bag were aggregated into a final variable set. Selected variables were validated against randomly resampled subsets to confirm stability of predictor selection. This strategy follows work showing that bagging approaches are effective for missing-data problems (Charlotte & Bello, 2023; Duangsoithong & Windeatt, 2010) and that imputation is not required with XGBoost (Ergul Aydin & Kamisli Ozturk, 2021). This final variable set of the most important 75 variables (15 x 5 bags) underwent the XGBoost feature selection algorithm under the same conditions as previous, to produce a final hierarchical ranking of predictors by their relative importance in the prediction of variance in SP.

Creation of Predictive Models Using Regression

Although XGBoost is effective for feature selection (Chen & Guestrin, 2016; Lundberg et al., 2020), it has limited interpretability, especially in understanding specific variance explained (Lipton, 2018). Therefore, regression was used on the remining 80% of the data to quantify the variance explained by the key predictors identified, adjusting for age and sex.

Imputation was considered before the regression analysis but was rejected on the basis that the missingness is not random and combining that with the heterogeneous and complex nature of the dataset, meant that there was a likelihood of this "missingness at random" assumption being violated and therefore introducing bias to the dataset. Instead, complete-case analysis was used to ensure that all parameter estimates were based solely on observed data. Although this approach reduced the sample size, the number of complete observations remained sufficiently large to maintain statistical power. This decision prioritised analytical transparency, interpretability, and reproducibility over the uncertain assumptions required for multiple imputation.

All candidate variables within the final variable set were initially analysed in individual linear regression models adjusted for age and sex. These models were used to rank predictors based on their covaried cross-validated coefficient of determination (R^2). Variables that were significantly auto correlated with the outcome variable (e.g. frequency of activities) were eliminated.

Variables were sequentially added to multivariable regression models in order of predictive strength. The sample size decreased as the number of variables included increased, as only participants with complete data on all the included variables were retained in the model. Model performance was assessed using 5-fold cross-validation. Upon completion of model selection, two summary models were constructed. The first model incorporated all predictors up to an elbow on the R^2 and RMSE curves, omitting variables that were not statistically significance ($p>0.05$) or had negligible contribution ($\beta<0.05$) to model performance, in order to optimise predictive accuracy and the proportion of variance explained. The second model applied additional filtering based on inter-variable correlations and individual predictor impact, yielding a parsimonious model to limit overfitting and improve explainability (Forster, 2000; Yarkoni & Westfall, 2017), while maintaining acceptable levels of model performance.

Development of Thematic Factors

Confirmatory Factor Analysis (CFA) was run on the primary theme identified using the *lavaan* package in R (Rosseel, 2012) to evaluate all possible subsets of 3–7 variables on an independent data sample. Each subset was tested on complete cases to avoid bias from missing data (Enders, 2010). For each CFA, standardized factor scores were extracted and regressed with social participation while adjusting for age and sex. Fit indices (CFI, TLI) were recorded and the proportion of variance explained (R^2) by the latent factor alone and with covariates. Subsets with Tucker-Lewis Index (TLI)<0.95 were excluded to ensure both adequate identification and model fit. No fixed loading threshold was applied at this point as the latent construct is potentially broad and the data sample is large - contemporary practice and recent methodological work emphasize interpreting loadings in context (sample size, fit indices, communalities, theoretical coverage) rather than using rigid cutoffs, and similar pragmatic decisions have been made in recent psychometric and applied research contexts(Cao et al., 2021; Flores-Kanter et al., 2021; Husain et al., 2025). This approach follows a confirmatory, theory-driven model selection framework that balances statistical fit with theoretical interpretability (Brown, 2015; Hu & Bentler, 1999).

Comparison of Predictive Models and Factors in Psychosis Relative to the Overall Population

Moderation analysis was completed on these predictive models and latent factor generated. Moderation was tested using the interaction term (Predictor x Psychosis) in the regression:

$$SP \sim Genetic\ Sex + Age + Predictor + Psychosis + (Predictor\ x\ Psychosis).$$

This approach tested whether the associations found between predictors and SP differed between populations with and without psychosis, while accounting for baseline group differences.

Chapter 3: Results

Participants

A total of 492,875 participants had available social participation (SP) data (54.3% female; mean age = 56.5±8.1 years). The mean SP score was 4.23 (female mean=4.4; male mean=4.0). Of these, 3,431 participants met criteria for psychosis. The breakdown of these statistics by presence of psychosis is given in Table 1, showing that the psychosis cohort are younger, have a higher percentage of males and lower social participation than the general population.

Table 1. Description of Overall Study Cohort including Comparison of the Population and Psychosis Subgroup

	Overall	General Population	Psychosis	Significance
Sample (n)	492,875	489,444	3,431	
Age (mean years ± SD)	56.5±8.1	56.6±8.1	55.8±8.4	Mann-Whitney U=8.8e08, p=1.1e-07
Sex (% Female)	54.3	54.3	51.8	Chi-Squared=7.85, p=0.005
Social Participation (Mean score ± SD)	4.23±1.49	4.23±1.49	4.10±1.72	Mann-Whitney U=8.6e08, p=0.002

Table 1 Legend: Sample characteristics stratified by psychosis status. Values represent mean ± SD or %. Tests of group differences between the general population and the psychosis cohort used Mann–Whitney U or chi-squared as appropriate.

Identification of Most Important Variables

An initial 329 psychosocial, lifestyle, and demographic predictors were identified. After preprocessing, decoding, and transformation, this increased to 514 numerical variables. Exclusion of variables with high collinearity (r>0.85) or missingness greater than 70% reduced this to 450 variables (Appendix 1 Table S1.2). Identification of the most important of these variables using XGBoost involved five subsets as individual bags (Table 2) with the top 15 predictors from each bag (Appendix 2) retained for establishing the optimal predictive models.

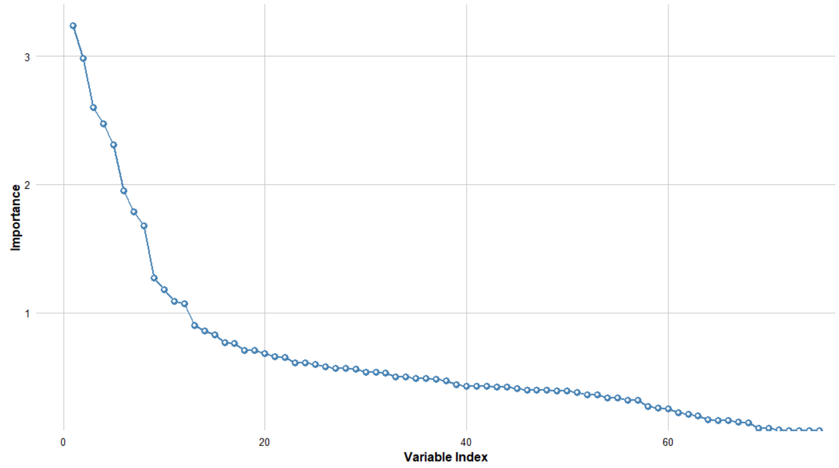
Table 2. Number of Variables in Each Bag, Based on Missingness

NA Percentage range	Bag ID	Number of variables
NA%=0	1	112
0%<NA% ≤ 10%	2	80
10%<NA% ≤ 15%	3	75
15%<NA% ≤ 65%	4	91
65%<NA% ≤ 70%	5	92

Table 2 Legend: Distribution of variables by missingness bag, using 20% of the data (NA = data not available).

The XGBoost model of the combined important parameters demonstrated that the most important predictors of SP included frequency and duration of exercise, ability to confide, employment status, alcohol consumption, and satisfaction with friendships and family relationships (Figure 5).

Figure 5. Relative Importance of Top Predictors of SP from XGBoost analysis



XGBoost-determined importance ranking of the top 20 predictors of Variance in Social Participation	Bag ID	Importance
Duration of other exercises	4	3.24
Frequency of other exercises in last 4 weeks	4	2.98
Length of working week for main job	4	2.60
Average weekly beer plus cider intake	4	2.47
Able to confide	2	2.31
Age at first live birth	4	1.95
Employment-retired	1	1.79
Number of children fathered	4	1.68
Place of birth in UK - east co ordinate	3	1.27
Friendships satisfaction	5	1.18
Number in household	1	1.09
Place of birth in UK - north co ordinate	3	1.07
Number of days week of vigorous physical activity 10 minutes	2	0.90
Frequency of travelling from home to job workplace	4	0.86
Mother's age	4	0.83
transport_leisure-motor car	1	0.77
Weekly usage of mobile phone in last 3 months	4	0.76
Length of time at current address	1	0.71
Activity-other exercises	1	0.71
Time spent watching television	1	0.68

Figure 5 Legend: The value of the overall importance of the most important variables where the input variables were the combination of the most important 15 variables from each of the initial bags, plotted in order of contribution. This shows that many variables contribute to SP (Appendix 2 Table S2.6 for full list). The importance value of the top 20 variables in the XGBoost analysis are shown in the panel. While Bag 5 variables ranked lower overall, several from Bag 4 scored highly, indicating that predictor importance reflected genuine associations rather than missingness.

Establishing the Predictive Models of Social Participation Using Regression

These top 75 predictors were evaluated as to their level of association with SP through regression models on the independent sample. The regression analysis indicated that many variables contributed to social participation (SP). Model performance improved as predictors were added, with root mean square error (RMSE) decreasing up to ~50 variables, after which overfitting was evident as RMSE increased (Hastie et al., 2001b). At this point, the cross-validated adjusted R^2 was ~0.135. An inflection point was observed at 20 predictors, yielding an adjusted R^2 of 0.1176, of which 0.0305 was explained by age and sex (Table 3).

Table 3. Cumulative Regression Results Showing the Variance of Social Participation Explained

Predictors	Cross Validated R^2	Cross Validated RMSE	P-Value	Number of Cases	ΔR^2
Genetic sex	0.0149	0.9917	<2.2e-16	385985	
+ Age at recruitment	0.0305	0.9838	<2.2e-16	385985	0.0156
+ Friendships satisfaction	0.0671	0.9652	<2.2e-16	130467	0.0366
+ Belief that own life is meaningful	0.0808	0.9513	<2.2e-16	43893	0.0137
+ Able to confide	0.0847	0.9501	<2.2e-16	42866	0.0039
+ Happiness	0.0845	0.95	0.0857	42771	-0.0002
+ Family relationship satisfaction	0.0837	0.9486	0.6891	42527	-0.0008
+ General happiness	0.0837	0.9486	0.0576	42372	0
+ employment-retired	0.098	0.9412	<2.2e-16	42372	0.0143
+ Felt loved as a child	0.0988	0.9403	<2.2e-16	42258	0.0008
+ Been in a confiding relationship	0.1005	0.9392	<2.2e-16	41272	0.0017
+ Alcohol intake frequency	0.1036	0.9375	<2.2e-16	41262	0.0031
+ Fed up feelings	0.1051	0.9371	<2.2e-16	40747	0.0015
+ transport_leisure-motor	0.1104	0.9342	<2.2e-16	40691	0.0053
+ Frequency of unenthusiasm disinterest	0.1104	0.9335	0.856	40097	0
+ Financial situation satisfaction	0.1103	0.9334	0.7196	40036	-0.0001
+ Neuroticism score	0.1109	0.933	0.2954	34949	0.0006
+ Health satisfaction	0.1109	0.9329	0.2646	34916	0
+ Length of mobile phone use	0.1148	0.9306	<2.2e-16	34779	0.0039
+ employment-employed	0.1176	0.9291	<2.2e-16	34779	0.0028

Table 3 Legend: Table showing the cumulative regression results. Predictors are added sequentially in order of incremental R^2 contribution after age and sex (Appendix 3 Table S3.1). Columns show cross validated R^2 and root mean square error (RMSE), significance, sample size, and change in R^2 from the addition of that variable (ΔR^2).

The initial list of predictors, based on these 20 variables, was reduced by excluding predictors that were not statistically significant ($p>0.05$) or small effect size ($\beta<0.05$). This yielded a model with 8 variables that explained 11.49% of the variance in SP (adjusted $R^2=0.1149$; Table 4). This reduction in variables resulted in a negligible decrease in explained variance and was selected as the full predictive model.

Table 4. Summary of Predictive Model of Social Participation

Predictor	Estimate (β)	Std. Error	t-value	P-value
(Intercept)	-0.0219	0.0047	-4.6851	<2e-16
Genetic sex	-0.1132	0.0047	-24.0383	<2e-16
Age at recruitment	0.0453	0.0062	7.3369	<2e-16
Friendships satisfaction	-0.1571	0.0049	-31.9079	<2e-16
Belief that own life is	0.0622	0.0047	13.1165	<2e-16
Able to confide	0.0603	0.005	11.9732	<2e-16
Employment-retired	0.0905	0.0083	10.9399	<2e-16
Alcohol intake frequency	-0.0641	0.0048	-13.2848	<2e-16
transport_leisure-motor	0.0769	0.0046	16.8251	<2e-16
Length of mobile phone use	0.0589	0.0048	12.2544	<2e-16
Employment-employed	-0.0871	0.0075	-11.6271	<2e-16

Table 4 Legend: Coefficients, standard errors, and significance levels for predictors retained in the reduced regression model. Cross validated R^2 0.1149 with cross validated RMSE of 0.9337 on 42,564 participants. SP was higher for females, older people, those satisfied with friendships, those who believe their lives are meaningful, those who have someone to confide in, those retired, those who drink more frequently, those who have access to a car for leisure activities, those who have been using a mobile phone for a long time and those who are not employed.

To derive a more parsimonious model, a stricter inclusion threshold was applied ($|t| \geq 18$). This resulted in a 4-predictor model that explained 10.16% of the variance in SP (adjusted R^2=0.1013; Table 5). Age and sex accounted for 3.78% of this variance. This parsimonious model demonstrates that the strongest predictors were satisfaction with friendships, belief that one's life is meaningful, employment status, and access to a motor vehicle.

Table 5. Summary of Parsimonious Predictive Model of SP

Predictor	Beta	Std. Error	t-value	P-Value
(Intercept)	-0.0076	0.0045	-1.6822	0.0925
Genetic sex	-0.1039	0.0046	-22.7392	<2e-16
Age at recruitment	0.0391	0.0061	6.4324	<2e-16
Friendships satisfaction	-0.1743	0.0048	-36.4199	<2e-16
Belief that own life is meaningful	0.0687	0.0047	14.7618	<2e-16
Employment-retired	0.1555	0.0059	26.5002	<2e-16
transport_leisure-motor	0.0804	0.0045	17.792	<2e-16

Table 5 Legend: Coefficients, standard errors, and significance levels for predictors retained in the 4-variable parsimonious model. Cross validated R^2 0.1013 with cross validated RMSE of 0.9404 on 43,832 participants.

An analysis was completed to examine whether the sample of data used in the complete cases for the regression analysis was representative of the overall sample or whether there were any significant differences due to the non-random missingness of data. Results are presented below in Table 6, showing that the complete cases data sample had identical SP to the full regression data sample, at 4.23. There were small but statistically significant differences in age (56.55 to 56.40) and presence of psychosis (0.696% to 0.606%) and a more significant difference in the sex distribution (% Female 54.3% to 57.1%)

Table 6. Description of Overall Study Cohort versus Complete Cases in Regression Models

	Regression Data Sample	Regression Complete Cases	Significance
Sample (n)	394,300	42,564	
Age (mean years ± SD)	56.55±8.1	56.40±7.8	t=-3.76, p=0.00017
Sex (% Female)	54.3	57.1	Chi-Squared=119.13, p<2.23-16
Social Participation (Mean score ± SD)	4.23±1.49	4.23±1.48	t=-0.319, p=0.75
% with Psychosis	0.696%	0.606%	Chi-Squared=4.43, p=0.035

Table 6 Legend: Comparison of complete cases sample with full data sample, showing matched SP, small differences in age and presence of psychosis, but a clear difference in genetic sex.

Development of Thematic Factor

The next step was to conduct confirmatory factor analysis (CFA) to identify whether a viable latent construct could be created from the data. From the initial analysis it was clear that the strongest predictors consistently clustered around emotional and relational wellbeing.

The latent construct was therefore modelled on an independent data sample using the 14 observed variables related to the theme of emotional wellbeing, that had been identified as important by the prior analysis (Appendix 4). The confirmatory, theory-driven model selection framework applied led to the development of a single 4-factor latent factor, comprising of friendship satisfaction, feeling loved as a child, are able to confide and the presence of fed-up feelings.

This latent factor (Figure 6) was selected based on the best fit indices (Comparative fit index, CFI=0.995, Tucker-Lewis Index, TLI=0.985) and explained more variance in SP than any of the individual variables (R^2=0.0835, with 0.0370 explained by age/sex).

Figure 6. Confirmatory Factor Analysis (CFA) for Emotional and Relational Wellbeing Latent Factor

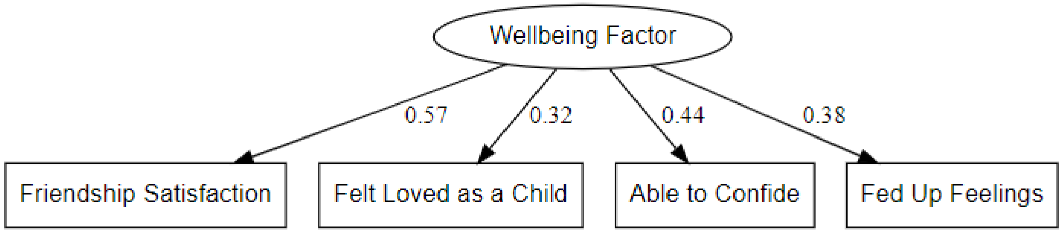


Figure 6 Legend: Emotional wellbeing factor, showing the loadings of the variables on to the latent factor. Fit indices were strong: CFI = 0.995, TLI = 0.985

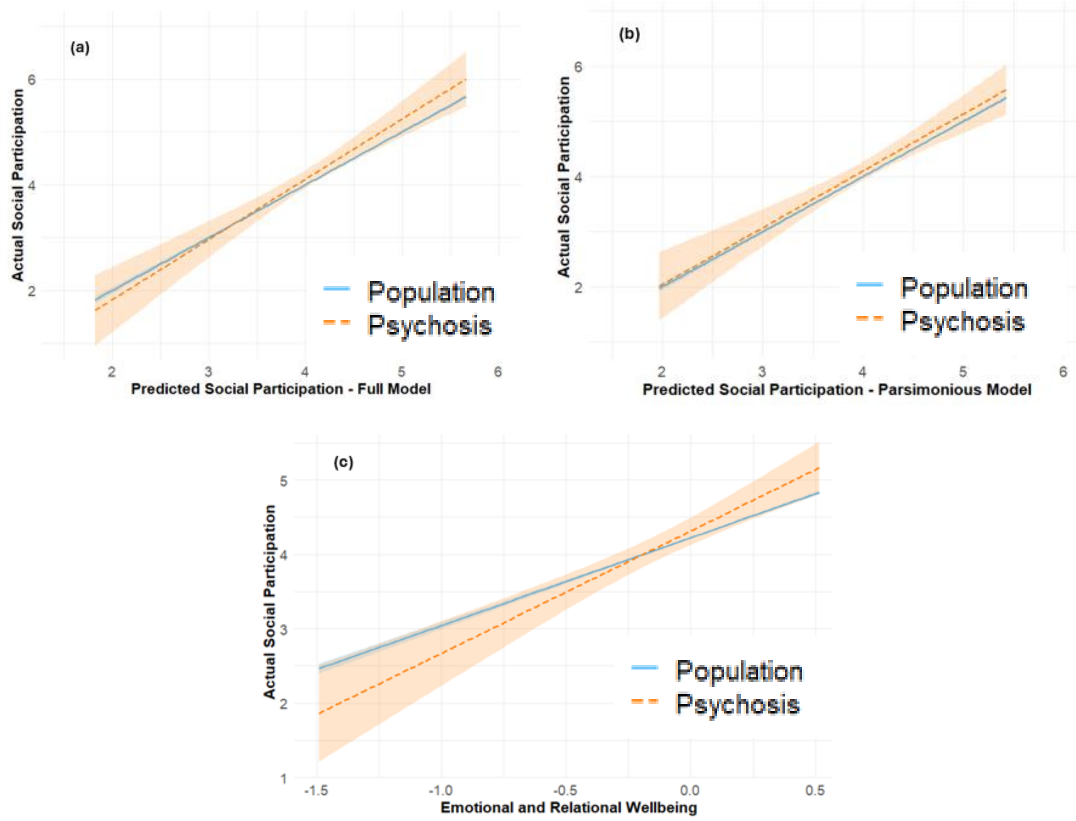
Comparison of Models in General Population and Psychosis Cohort

The analysis to determine whether the association between the predictive models and latent factors and social participation was moderated by the presence of psychosis used the samples described in Table 1.

These models and factors all demonstrate a strong and significant effect on social participation for those with and without psychosis (Figure 7). None of the interaction terms (predictor*psychosis) reached statistical significance (all $p>0.05$), suggesting that psychosis status does not significantly moderate the relationship between any of the predictors and social participation. Although confidence intervals for the psychosis group were wider due to smaller sample size, the slopes were similar to those of the population cohort. This indicated that the associations observed were consistent across psychosis and non-psychosis cohorts, and that emotional and relational wellbeing remained a valid predictor of social participation even in the presence of psychosis.

The interaction plot shows the similarity in the relationships between predicted and actual social participation, for those with and without psychosis. Similarly, the relationship of the latent wellbeing factor with social participation does not differ statistically significantly between those groups with or without psychosis (Appendix 5 Table S5.1).

Figure 7. Predictive Relationships and Latent Variable Do Not Significantly Differ between the General Population and the Psychosis Group



Predictor	Interaction Term	F-stat	Interaction p-value	R^2	Effect Size(d) Population (N)	Effect Size(d) Psychosis (N)	Main Effect(β) Psychosis
SP Full model	Full model * Psychosis	0.85	0.358	0.115	1.00*** (42306)	1.14*** (258)	0.137
SP Parsimonious model	Parsimonious model * Psychosis	0.04	0.839	0.102	1.00*** (43567)	1.03*** (265)	0.110
Latent – Wellbeing Factor	Wellbeing Factor * Psychosis	2.38	0.123	0.084	1.083*** (43932)	1.459*** (274)	0.124

Figure 7 Legend: Moderation effect estimate for the presence of psychosis on the relationship between the predictive models and the latent wellbeing measure and social participation. Effect sizes (Cohen's *d*) are reported separately for the general population and psychosis groups, while the main effect of psychosis is represented by standardized regression coefficients (β). Sample sizes for each analysis are shown in parentheses. Across the (a) full and (b) parsimonious predictive models, and (c) the latent wellbeing factor, significant positive associations were observed between the predictors and social participation. However, no predictor × psychosis interaction terms reached significance, indicating that these relationships did not differ significantly between individuals with and without psychosis. ***p < .001.

Chapter 4: Discussion

Predictability of Social Participation

Analysis of over 500 aspects of life, our environment, health and wellbeing showed that numerous factors are related to social participation and that the strongest statistical predictor is emotional and relational or intrapersonal wellbeing. The largest gains in variance-explained occurred in the first 20 predictors (Table 3). Among these, the most influential were satisfaction with friendships, belief that one's own life is meaningful, the ability to confide in others, and feeling loved during childhood. These factors consistently appeared as dominant contributors to variance in adult self-reported social participation (SP), with friendship satisfaction alone accounting for an increase of 3.7% in the variance explained when accounting for age and sex. Together, these results suggest a significant association between individuals who feel connected, valued, and supported emotionally and engagement in socially participatory behaviours.

The four-variable parsimonious model (SP higher if satisfied with friends, belief that life is meaningful, have access to private transport and are retired) explains 10.16% of the variance in social participation, only slightly less than the full model. This small reduction in explanatory power underscores the centrality of these few factors in predicting social participation. The pragmatic implications are considerable; interventions aimed at improving perceived life meaning or logistical access such as transportation for older or vulnerable adults, may substantially enhance social participation, without requiring broad-spectrum interventions (Lehning et al., 2018; Levasseur et al., 2015).

These results also speak to the principle of parsimony in population health modelling. While the temptation in large biobank datasets is to pursue ever-expanding models, the evidence here suggests that a small subset of psychosocial variables exerts disproportionate influence. This has important implications for intervention prioritisation: policy and service design may yield more impact by targeting a handful of modifiable factors rather than diluting resources across dozens of weakly predictive variables.

Emotional and Relational Wellbeing Construct

The latent factor representing emotional and relational wellbeing, which combined friendship satisfaction, belief one's own life is meaningful, having been in a confiding relationship as an adult and the presence of fed-up feelings, demonstrated a correlation of 8.35% with SP. This factor explained more variance in social participation than any single contributing variable and demonstrated excellent model fit, supporting its stability and validity as a construct (Sathyanarayana & Mohanasundaram, 2024) for representing emotional and relational wellbeing. The convergence of distinct but interrelated emotional experiences into a single predictive factor reinforces the theoretical framing of social participation as deeply embedded in subjective wellbeing and internalised social support.

This latent construct aligns closely with theories of social capital, which emphasise the value of interpersonal trust and reciprocity in enabling participation, as well as with self-determination theory, which highlights relatedness as a basic psychological need. The coherence of the factor thus suggests that social participation may be best understood not as an isolated behavioural outcome, but as an emergent property of broader emotional and relational wellbeing. In turn, this has implications for intervention: strengthening relational quality and meaning-making may indirectly foster participation more effectively than attempting to directly stimulate social behaviours in isolation.

Models in Psychosis Cohort

The next step in this analysis was to explore whether the presence of psychosis might be associated with an altered relationship between the predictor variables and social participation. Across all results, including the full 8-parameter regression, the more streamlined 4-parameter version, and analyses using the latent factor, the patterns of association with social participation appeared consistent regardless of psychosis status. Interaction terms involving psychosis were uniformly non-significant, suggesting that the presence or absence of psychosis did not alter the pattern of associations with social participation and that the same emotional wellbeing factors underpin social participation regardless of psychiatric status. This underscores the relevance of psychosocial and wellbeing-focused interventions in psychosis, complementing pharmacological and other interventions. This finding reinforces research (Gagiu et al., 2024; Phillips, 1967) highlighting that perceived life satisfaction and social-emotional variables maintain relevance for social engagement even in the context of severe psychiatric disorder. There is also cross-sectional evidence (Degnan et al., 2018) that larger social network size may improve outcomes in schizophrenia, supporting the importance of psychosocial interventions in psychiatric populations.

The lack of significant moderation by psychosis is notable. It suggests that barriers to participation in psychosis may not lie in different determinants, but rather in the amplification of challenges such as stigma, service access, and illness burden. From a translational standpoint, this points toward scaling interventions that target relational wellbeing across populations, while tailoring delivery to address illness-specific barriers. Such a conclusion resonates with contemporary recovery models, which emphasise personal meaning and social inclusion as central to recovery regardless of diagnostic category.

Comparison with Literature

Although our models explained up to 13.5% of the variance in SP, the majority remains unexplained. This highlights the complexity of SP, which is likely also shaped by unmeasured contextual and cultural factors, as well as the limitations of available measures within UK Biobank. This finding aligns with broader literature in population-level mental health and social psychiatry, where measured psychosocial, genetic or brain variables do not account for the majority of the variance in social participation (Casburn et al., 2025; Doherty et al., 2025; Wilding et al., 2023). The limited incremental value added by more than 18 predictors further supports the law of diminishing returns in complex psychological modelling (Yarkoni & Westfall, 2017) and justifies a shift toward latent variable-based interpretation and intervention.

These findings broadly align with prior studies that link emotional wellbeing with community engagement (Umberson & Montez, 2010; Wickramaratne et al., 2022). Other studies have also emphasised deprivation indices such as income, housing tenure and the Townsend Deprivation Index as predictors of SP (Ashida et al., 2016; Kung et al., 2022; P. Townsend, 1987). An additional insight that this analysis has provided is that financial satisfaction emerged as more predictive than objective measures of actual income or deprivation, highlighting that perceived financial security may shape opportunities and motivation for social engagement more directly than material indicators alone.

Employment status, a key socioeconomic indicator, and a component of most deprivation indices, also showed a meaningful relationship with social participation. Being retired, in particular, was associated with greater social engagement, while being employed was negatively associated. Retired individuals may experience both greater temporal freedom and contentment with their financial situation, fostering social activity, while those in active employment may face time-poverty or role strain that constrains participation. Thus, while broad deprivation indices were not the most strongly predictive in this sample, specific aspects of socioeconomic status such as access to motor transport, subjective financial satisfaction and employment status emerged as specifically relevant.

Several psychosocial variables are frequently associated with social participation in prior research but did not influence SP very strongly in this analysis. Measures of childhood trauma or other adverse early life events, did not remain significant in the final models. Although childhood adversity has been linked to long-term social outcomes (Reinhard et al., 2022), its incremental predictive value here was minimal once current wellbeing variables were included. The UK Biobank cohort is generally healthier, wealthier, and older (mean age 56 years) than the wider UK population which may mean that the Biobank sample, recruited in midlife and later adulthood, may underrepresent younger individuals for whom early adversity exerts stronger immediate impacts on participation. Likewise, cultural and generational shifts in attitudes toward participation and wellbeing may alter predictor importance across cohorts. These findings contrast with some prior research highlighting the long-term social consequences of childhood trauma (Allen et al., 2023; Hughes et al., 2017; Jenkins et al., 2020). However, these results align with suggestions that early adversity effects are mediated by emotional wellbeing (Kobrinsky & Siedlecki, 2022; Wu et al., 2025).

Overall, the comparison with the literature illustrates the importance of cohort characteristics, measurement specificity, and analytic strategy in shaping conclusions about social participation. A particular strength of this study is the application of machine learning methods (XGBoost) in a population-scale dataset, which allows both breadth of predictors and robustness of variable selection. However, this strength is balanced by the limits of cross-sectional design, the UK Biobank's well-documented selection biases, and the challenge of translating statistical importance into actionable causal inference.

Nevertheless, the convergence of emotional wellbeing variables across regression and factor analysis, and their stability across healthy populations and those with psychosis, suggests these predictors may serve as viable targets for intervention. The model's robustness supports future use in stratified or targeted interventions and the reproducibility of findings across models with different complexity levels enhances confidence in the validity of these constructs. Furthermore, the identification of practical markers such as transport access and employment status, despite their lower impact to social participation, adds utility for intervention design, especially in community and policy contexts.

Limitations

The study's conclusions must be interpreted in light of several limitations. Most importantly, substantial data loss occurred due to missing values across predictors, resulting in reduced sample size for key analyses. Whilst the data sample retained significant power in the regression analysis, this data reduction does limit statistical power to detect moderation effects due to the reduction in of numbers in the psychosis subgroup. As a result, null findings in the interaction terms should be interpreted cautiously, as they may reflect insufficient power rather than genuine absence of group differences.

The cross-sectional design also constrains causal inference: emotional wellbeing may promote SP, SP may enhance wellbeing, or both may be influenced by unmeasured third variables. Future longitudinal analyses, using repeated UK Biobank measures or other cohort studies, will be essential to clarify causal directions.

There are limitations based on the UK BioBank data sample. Firstly, our definition of social participation was constrained by available measures and may not capture the full construct, although the UKB SP measures have been used successfully previously (Foster et al., 2023; O'Connell et al., 2024). Another limitation involves the representativeness of the UK Biobank cohort, which underrepresents younger, lower-income, and less healthy individuals. The subsample for the regression analysis, that had the additional requirement of having a complete data set, also had some statistical differences to the overall data set, so this subsample does not meet the criterion of missingness being at random. It should especially be noted that there is less complete data for those with psychosis than those in the general population. The fact that this subsample, and the UK BioBank in general, are not fully representative of the general UK population, combined with the fact that many of the predictors vary across cohorts is a limitation that needs to be considered in generalising these findings, especially considering that SP increases with age, health and wealth.

Despite these limitations, the consistency of findings across models supports the robustness of the core associations identified and provides a foundation for targeted, testable interventions in future research.

Summary

In summary, life satisfaction and emotional connectedness emerged as the most robust and generalisable correlates of social participation across a large population-based sample, including those with psychosis. The findings demonstrate that, although demographic and lifestyle variables contribute modestly, it is subjective measures of wellbeing, particularly those related to relationships, meaning, and emotional security, that most consistently explain variance in SP. These associations persist even in individuals with severe psychiatric illness, indicating that the same foundational factors underpin social engagement across the mental health spectrum. Therefore, it appears that interventions aimed at fostering subjective wellbeing, emotional resilience, and perceived social value may offer a viable pathway to improving social integration and reducing associated health burdens in both clinical and general populations.

Taken together, these findings support a conceptualisation of social participation as primarily rooted in emotional and relational domains, with socioeconomic and lifestyle variables playing important but secondary roles. This framing opens a pathway for developing holistic intervention models that bridge mental health care, social policy, and community design, emphasising both structural supports (e.g., transport, retirement policies) and intrapersonal strengths (e.g., resilience, meaning-making). Such an approach would align public health with recovery-oriented practice, ensuring that strategies to enhance social participation are relevant across the spectrum from general population wellbeing to severe psychiatric illness.

Appendices:

Appendix 1

Table S1.1: Showcase data fields selected for variable analysis:

- Sociodemographics
    - Household
    - Employment
    - Education
    - Ethnicity
    - Other
- Lifestyle and Environment
    - Physical Activity
    - Electronic Device Use
    - Sleep
    - Alcohol
    - Sexual Factors
    - Early Life Factors
    - Family History
- Psychosocial Factors
    - Social Supports
    - Mental Health
- Health and Medical History
    - Eyesight
    - Mouth
    - General Health
    - Breathing
    - Artery Disease
    - Pain
    - Cancer screening
    - Operations
    - Medical Conditions
    - Medication
    - Hearing
    - ICD-10 Diagnoses
- Verbal Interview
    - Early life factors
    - Employment
    - Medical Conditions
    - Medications
    - Operations
- Mental Wellbeing
    - Adverse Life Events

    - Anxiety and Panic
    - Mood changes
- Mental Health
    - Depression
    - Unusual and Psychotic Experiences
    - Traumatic Events
    - Self-harm behaviours
    - Happiness and subjective wellbeing
- Local Environment
    - Home locations
    - Pollution – air and noise
    - Greenspace and Coastal Proximity
    - Water Quality

Table S1.2: Detailed List of Variables Investigated:

Eid – Unique Identifier	Distance between home and job workplace
Edited_Social – Social Participation Variable	Job involves mainly walking or standing
Frequency of light DIY in last 4 weeks	Job involves heavy manual or physical work
Duration of light DIY	Job involves shift work
Time spent watching television TV	Age completed full time education
Time spent using computer	Number of days week walked 10 minutes
Time spent driving	Duration of walks
Drive faster than motorway speed limit	Number of days week of moderate physical activity 10 minutes
Length of mobile phone use	Duration of moderate activity
Weekly usage of mobile phone in last 3 months	Number of days week of vigorous physical activity 10 minutes
Hands free device speakerphone use with mobile phone in last 3 month	Duration of vigorous activity
Difference in mobile phone use compared to two years previously	Usual walking pace
Usual side of head for mobile phone use	Frequency of stair climbing in last 4 weeks
Sleep duration	Frequency of walking for pleasure in last 4 weeks
Getting up in morning	Duration walking for pleasure
Morning evening person chronotype	Depression-talk therapies
Nap during day	Depression-other activities
Birth weight known	attendance_allowance
Sleeplessness insomnia	disability_allowance
Mobility problems today	bluebadge_allowance
Self care problems today	DVT
Pain discomfort today	Emphysema
Anxiety depression today	lung_bloodclot
Scale to indicate how health is today	Asthma
Snoring	Hayfever/eczema
Daytime dozing sleeping	Cancer code-breast cancer
Place of birth in UK north co ordinate	Cancer code-basal cell carcinoma
Place of birth in UK east co ordinate	Cancer code - malignant melanoma
Number of operations self reported	Cancer code-prostate cancer
Alcohol intake frequency	Cancer code-cervical cancer
Average weekly red wine intake	Cancer code-precervical cancer
Average weekly champagne plus white wine intake	Cancer code-colon cancer
Average weekly beer plus cider intake	Cancer code-skin cancer
Average weekly spirits intake	Cancer code- uterine cancer
Average weekly fortified wine intake	Cancer code - bladder cancer
Alcohol usually taken with meals	Employment-employed
Alcohol intake versus 10 years previously	Employment-retired
Country of birth UK elsewhere	Employment-homemaker
Breastfed as a baby	Employment-unable_sick_disabled
Comparative body size at age 10	Employment-unemployed

Appendices:

Comparative height size at age 10
Handedness chirality laterality
Adopted as a child
Part of a multiple birth
Maternal smoking around birth
Father still alive
Father s age at death
Mother still alive
Mother s age
Number of full brothers
Number of full sisters
Mood swings
Miserableness
Irritability
Sensitivity   hurt feelings
Fed up feelings
Nervous feelings
Worrier   anxious feelings
Tense   highly strung
Worry too long after embarrassment
Birth weight
Alcohol drinker status
Neuroticism score
Loneliness isolation
Job code at visit
Guilty feelings
Risk taking
Ever addicted to any substance or behaviour
Ever felt worried tense or anxious for most of a month or longer
Ever worried more than most people would in similar situation
Ever had prolonged loss of interest in normal activities
Ever had prolonged feelings of sadness or depression
Ever taken cannabis
General happiness
General happiness with own health
Belief that own life is meaningful
Ever heard an unreal voice
Ever believed in an unreal conspiracy against self
Ever seen an unreal vision
Ever believed in unreal communications or signs
Ever thought that life not worth living
Ever self harmed
Ever contemplated self harm
Felt hated by family member as a child
Physically abused by family as a child
Felt loved as a child
Sexually molested as a child
Someone to take to doctor when needed as a child
Avoided activities or situations because of previous stressful experience in past month
Repeated disturbing thoughts of stressful experience in past month
Felt very upset when reminded of stressful experience in past month
Ever sought or received professional help for mental distress
Frequency of depressed mood in last 2 weeks
Ever suffered mental distress preventing usual activities
Ever had period of mania   excitability
Ever had period extreme irritability
Recent easy annoyance or irritability
Recent feelings or nervousness or anxiety
Recent feelings of inadequacy
Recent trouble concentrating on things
Recent inability to stop or control worrying
Recent feelings of depression
Recent poor appetite or overeating
Recent feelings of foreboding

Employment-unpaid_voluntary
Employment-student
Employment-other
Diagnoses main ICD10_poor_vision
Diagnoses main ICD10_incontinence
Diagnoses main ICD10_amnesia
Diagnoses main ICD10_stuttering
Diagnoses main ICD10_chestpain
Diagnoses main ICD10_cataract
Diagnoses main ICD10_diverticulardiseaselargeintestine
Diagnoses main ICD10_atherosclerotic_heartdisease
Diagnoses main ICD10_morbidity
Diagnoses main ICD10_inguinalhernia
Diagnoses main ICD10_gonarthrosis
Diagnoses main ICD10_colonpolyp
Diagnoses main ICD10_abdominalpain
Diagnoses main ICD10_senilecataract
Diagnoses main ICD10_haematuria
Diagnoses main ICD10_diaphragmatic_hernia
Diagnoses main ICD10_coxarthosis
Diagnoses main ICD10_uti
Diagnoses main ICD10_gastritis
Diagnoses main ICD10_gastroenteritis_colitis
Diagnoses main ICD10_changeonbowelhabit
Diagnoses main ICD10_carpaltunnelsyndrome
Diagnoses main ICD10_upperadomenpain
Diagnoses main ICD10_analhaemorrhage
Diagnoses main ICD10_oesophagitis_reflux
Diagnoses main ICD10_intestinalneoplasmscreening
Diagnoses main ICD10_collapsing
Diagnoses main ICD10_malignantmelanoma_face
Diagnoses main ICD10_varicoseveins_lower_ext
Diagnoses main ICD10_dyspepsia
Diagnoses main ICD10_malignantneoplasm_prostate
Diagnoses main ICD10_lobarpneumonia
Diagnoses main ICD10_malignantneoplasm_breast
Diagnoses main ICD10_hyperplasia_prostate
Diagnoses main ICD10_anaemia
Diagnoses main ICD10_reflux
Diagnoses main ICD10_atrial_fibrillation
Diagnoses main ICD10_anaemia_iron_deficiency
Diagnoses main ICD10_gallbladder_calculus
Diagnoses main ICD10_lower_respitory_infection
Diagnoses main ICD10_gastroentiritis
Diagnoses main ICD10_postmenopausal_bleeding
Diagnoses main ICD10_constipation
eye_diabetes
eye_glaucoma
eye_trauma
eye_cataract
eye_macular
eye_other
fracture_ankle
fracture_leg
fracture_hip
fracture_spine
fracture_wrist
fracture_arm
fracture_other
Fuel_gas_cooker
Fuel_gas_fire
Fuel_open_fire
Fuel_other
heating_other

Appendices:

Recent thoughts of suicide or self harm
Recent lack of interest or pleasure in doing things
Recent trouble relaxing
Recent restlessness
Trouble falling or staying asleep or sleeping too much
Recent changes in speed amount of moving or speaking
Recent feelings of tiredness or low energy
Recent worrying too much about different things
Belittlement by partner or ex partner as an adult
Been in a confiding relationship as an adult
Physical violence by partner or ex partner as an adult
Sexual interference by partner or ex partner without consent as an adult
Able to pay rent mortgage as an adult
Been in serious accident believed to be life threatening
Been involved in combat or exposed to war zone
Diagnosed with life threatening illness
Victim of physically violent crime
Witnessed sudden violent death
Victim of sexual assault
Frequency of unenthusiasm   disinterest in last 2 weeks
Frequency of tenseness   restlessness in last 2 weeks
Frequency of tiredness   lethargy in last 2 weeks
Seen doctor GP for nerves anxiety tension or depression
Seen a psychiatrist for nerves anxiety tension or depression
Ethnic background
Age at recruitment
Able to confide
Water hardness USGS classification
Water hardness WHO classification
Ca concentration
Mg concentration
Answered sexual history questions
Age first had sexual intercourse
Lifetime number of sexual partners
Ever had same sex intercourse
Overall health rating
Genetic sex
IPAQ activity group
Summed days activity
Summed minutes activity
At or above moderate vigorous recommendation
MET minutes per week for walking
MET minutes per week for moderate activity
MET minutes per week for vigorous activity
Summed MET minutes per week for all activity
Wears glasses or contact lenses
Age started wearing glasses or contact lenses
Townsend deprivation index at recruitment
Other eye problems
Plays computer games
Hearing difficulty problems
Hearing difficulty problems with background noise
Falls in the last year
Weight change compared with 1 year ago
Wheeze or whistling in the chest in last year
Chest pain or discomfort
Ever had bowel cancer screening
Most recent bowel cancer screening
Ever had prostate specific antigen PSA test
Relative age of first facial hair
Relative age voice broke
Hair balding pattern
Nitrogen dioxide air pollution 2010
Nitrogen oxides air pollution 2010

heating_gas
heating_electric
heating_oil
heating_portable_gas_paraffin
heating_solidfuel_ch
heating_open_fire
household_relations_partner
household_relations_child
household_relations_sibling
household_relations_parent
household_relations_grandparent
household_relations_grandchild
household_relations_other_related
household_relations_other_unrelated
stress_serious_injury_self
stress_serious_injury_closerelative
stress_death_close_relative
stress_death_spouse
stress_divorce
stress_financial
illnessfather_heartdisease
illnessfather_stroke
illnessfather_highbloodpressure
illnessfather_emphysema
illnessfather_diabetes
illnessfather_lungcancer
illnessfather_prostatecancer
illnessfather_bowelcancer
illnessfather_dementia
illnessfather_severedepression
illnessfather_parkinsons
illnessmother_heartdisease
illnessmother_stroke
illnessmother_highbloodpressure
illnessmother_emphysema
illnessmother_diabetes
illnessmother_lungcancer
illnessmother_breastcancer
illnessmother_bowelcancer
illnessmother_dementia
illnessmother_severedepression
illnessmother_parkinsons
illnesssibling_heartdisease
illnesssibling_stroke
illnesssibling_highbloodpressure
illnesssibling_emphysema
illnesssibling_diabetes
illnesssibling_lungcancer
illnesssibling_prostatecancer
illnesssibling_bowelcancer
illnesssibling_dementia
illnesssibling_severedepression
illnesssibling_parkinsons
manic_active
manic_talkative
manic_lesssleep
manic_creative
manic_all
maniamanifestation_talkative
maniamanifestation_restless
maniamanifestation_racingthoughts
maniamanifestation_sleepless
maniamanifestation_creativeideas
maniamanifestation_distracted

Appendices:

Particulate matter air pollution pm10 2010
Particulate matter air pollution pm2 5 2010
Particulate matter air pollution pm2 5 absorbance 2010
Particulate matter air pollution 2 5 10um 2010
Traffic intensity on the nearest road
Inverse distance to the nearest road
Traffic intensity on the nearest major road
Inverse distance to the nearest major road
Total traffic load on major roads
Close to major road
Sum of road length of major roads within 100m
Average 24-hour sound level of noise pollution
Number of children fathered
Had major operations
Diabetes diagnosed by doctor
Domestic garden percentage buffer 1000m
Water percentage buffer 1000m
Greenspace percentage buffer 300m
Domestic garden percentage buffer 300m
Water percentage buffer 300m
Natural environment percentage buffer 1000m
Natural environment percentage buffer 300m
Distance Euclidean to coast
Cancer diagnosed by doctor
Other serious medical condition disability diagnosed by doctor
Taking other prescription medications
Frequency of heavy DIY in last 4 weeks
Duration of heavy DIY
Index of Multiple Deprivation England
Employment score England
Health score England
Education score England
Housing score England
Crime score England
Living environment score England
Reason for reducing amount of alcohol drunk
Ever had breast cancer screening mammogram
Years since last breast cancer screening mammogram
Ever had cervical smear test
Years since last cervical smear test
Age when periods started menarche
Had menopause
Number of live births
Birth weight of first child
Age at first live birth
Age at last live birth
Ever had stillbirth spontaneous miscarriage or termination
Ever taken oral contraceptive pill
Age started oral contraceptive pill
Age when last used oral contraceptive pill
Ever used hormone replacement therapy HRT
Bilateral oophorectomy both ovaries removed
Had other major operations
Pace maker
Pregnant
Hearing aid user
Mother s age at death
Age at menopause last menstrual period
Ever had hysterectomy womb removed
Frequency of other exercises in last 4 weeks
Duration of other exercises
Non accidental death in close genetic family
Happiness
Work job satisfaction

maniamanifestation_confident
maniamanifestation_active
medication__cholesterol
medication__bloodpressure
medication__insulin
medication_cholesterol
medication_bloodpressure
medication_insulin
medication_HRT
medication_oralcontraception
medication_aspirin
medication_ibuprofen
medication_paracetamol
medication_ranitidine
medication_omeprazole
medication_lazatives
MH_social_anxiety_phobia
MH_schizophrenia
MH_other_psychosis
MH_personality_disorder
MH_phobia_other
MH_panic_attacks
MH_OCD
MH_mania_bp_depression
MH_depression
MH_bulimia
MH_overeating
MH_ASD
MH_anxiety_disorder
MH_anorexia
MH_agoraphobia
MH_ADD_ADHD
dietarysupp_fishoil
dietarysupp_glucosamine
dietarysupp_calcium
dietarysupp_zinc
dietarysupp_iron
dietarysupp_selenium
mouth_ulcers
mouth_painfulgums
mouth_bleedinggums
mouth_looseteeth
mouth_toothache
mouth_dentures
pain_headache
pain_facial
pain_neck_shoulder
pain_back
pain_abdominal
pain_hip
pain_knee
pain_whole_body
shortsighted
longsighted
reading_glasses
astigmatism
squint
lazyeye
other_eye_condition
anxiety_subtance_drugs_alcohol
anxiety_subtance_unprescribedmedication
anxiety_subtance_prescribedmedication
depression_subtance_drugs_alcohol
depression_subtance_unprescribedmedication

Health satisfaction	depression_subtance_prescribedmedication
Family relationship satisfaction	transport_work_none
Friendships satisfaction	transport_work_motor_vehicle
Financial situation satisfaction	transport_work_walk
Ever depressed for a whole week	transport_work_public_transport
Ever unenthusiastic disinterested for a whole week	transport_work_cycle
Ever manic hyper for 2 days	transport_notwork_motor
Ever highly irritable argumentative for 2 days	transport_notwork_walk
Private healthcare	transport_notwork_publictransport
Shortness of breath walking on level ground	transport_notwork_cycle
Leg pain on walking	activity_walking
Cochlear implant	activity_otherexercises
Tinnitus	activity_strenuoussports
Noisy workplace	activity_lightDIY
Loud music exposure frequency	activity_heavyDIY
Number of older siblings	activity_none
Month of birth	heart_attack
Type of accommodation lived in	Angina
Own or rent accommodation lived in	Stroke
Length of time at current address	highbloodpressure
Number in household	medication_VitA
Number of vehicles in household	medication_VitB
Average total household income before tax	medication_VitC
Time employed in main current job	medication_VitD
Length of working week for main job	medication_VitE
Frequency of travelling from home to job workplace	EduYears

Table S1.3: Variables that required specific interpretation:

Time spent watching TV/computer/driving: -10 → 0.

Alcohol usually taken with meals: -6 → 0.5.

Age first had sexual intercourse: -2 → NA.

Length of time at current address: -10 → 0

Number of days/weeks walked 10 minutes: -2 → NA.

Appendix 2

Full Importance Tables from XGBoost Analysis

Table S2.1 - The most important variables that had no NA values

Variable	**Importance**
activity_otherexercises	9.449327
employment_retired	3.957746
Number.in.household...Instance.0	2.290716
Distance..Euclidean..to.coast...Instance.0	2.012816
Length.of.time.at.current.address...Instance.0	1.873073
employment_employed	1.804057
Townsend.deprivation.index.at.recruitment	1.684512
Alcohol.intake.frequency....Instance.0	1.645012
Age.at.recruitment	1.554369
Natural.environment.percentage..buffer.1000m...Instance.0	1.472724
transport_notwork_motor	1.388269
Time.spent.watching.television..TV....Instance.0	1.374909
activity_strenuoussports	1.235057
Time.spent.using.computer...Instance.0	1.068411
activity_walking	1.016041
Country.of.birth..UK.elsewhere....Instance.0	1.006983
Natural.environment.percentage..buffer.300m...Instance.0	0.930877
Frequency.of.stair.climbing.in.last.4.weeks...Instance.0	0.839735
Birth.weight.known...Instance.0	0.818772
Overall.health.rating...Instance.0	0.744172
Sleep.duration...Instance.0	0.731938
Number.of.operations..self.reported...Instance.0	0.707731
Water.hardness..USGS.classification....Instance.0	0.705436
activity_lightDIY	0.669759
Number.of.vehicles.in.household...Instance.0	0.64245
employment_homemaker	0.493239
transport_notwork_cycle	0.459995
stress_serious_injury_closerelative	0.45385
Month.of.birth	0.407386
Fuel_gas_fire	0.375015
Water.hardness..WHO.classification....Instance.0	0.337524
activity_none	0.322077
Getting.up.in.morning...Instance.0	0.30731
Falls.in.the.last.year...Instance.0	0.275527
Seen.doctor..GP..for.nerves..anxiety..tension.or.depression...Instance.0	0.226566
Plays.computer.games...Instance.0	0.220563
Nap.during.day...Instance.0	0.219613
Usual.walking.pace...Instance.0	0.209748
Daytime.dozing...sleeping...Instance.0	0.200852

Sleeplessness...insomnia...Instance.0	0.172579
Type.of.accommodation.lived.in...Instance.0	0.169984
stress_death_close_relative	0.169892
medication_paracetamol	0.160237
Cancer.diagnosed.by.doctor...Instance.0	0.15371
Taking.other.prescription.medications...Instance.0	0.134954
Handedness..chirality.laterality....Instance.0	0.131573
Seen.a.psychiatrist.for.nerves..anxiety..tension.or.depression...Instance.0	0.130681
dietarysupp_glucosamine	0.13053
transport_notwork_publictransport	0.120966
transport_notwork_walk	0.120587
medication_omeprazole	0.114565
Answered.sexual.history.questions...Instance.0	0.110992
Other.eye.problems...Instance.0	0.109372
Wears.glasses.or.contact.lenses...Instance.0	0.102744
Fuel_other	0.10247
Diabetes.diagnosed.by.doctor...Instance.0	0.096572
mouth_bleedinggums	0.091239
dietarysupp_fishoil	0.089419
medication_aspirin	0.086125
activity_heavyDIY	0.084045
asthma	0.07978
stress_financial	0.076442
stress_serious_injury_self	0.076134
pain_hip	0.075155
mouth_ulcers	0.072647
mouth_dentures	0.072499
Fuel_gas_cooker	0.071438
pain_knee	0.07079
pain_back	0.0689
highbloodpressure	0.067572
lung_bloodclot	0.067078
pain_headache	0.066061
pain_neck_shoulder	0.066054
medication_ibuprofen	0.06253
dietarysupp_calcium	0.061634
medication_VitC	0.060097
Fuel_open_fire	0.059528
hayfever_eczema	0.05629
pain_abdominal	0.051986

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from bag 1

Table S2.2 - The most important variables that had NA values 0% < NA% ≤10%

Variable	Importance

Number of days week of vigorous physical activity 10 minutes	6.34
Able to confide	4.40
Genetic sex	4.32
Own or rent accommodation lived in	3.42
Ca concentration	2.87
Number of days week of moderate physical activity 10 minutes	2.60
Length of mobile phone use	1.72
Number of days week walked 10 minutes	1.64
Fed up feelings	1.39
Time spent driving	1.27
Nitrogen dioxide air pollution 2010	1.15
Nitrogen oxides air pollution 2010	0.96
Frequency of unenthusiasm disinterest in last 2 weeks	0.95
Traffic intensity on the nearest major road	0.92
Average 24-hour sound level of noise pollution	0.85
Particulate matter air pollution pm10 2010	0.82
Alcohol intake versus 10 years previously	0.82
Particulate matter air pollution pm2 5 2010	0.80
Particulate matter air pollution pm2 5 absorbance 2010	0.79
Drive faster than motorway speed limit	0.79
Inverse distance to the nearest major road	0.69
EduYears	0.69
Ever had bowel cancer screening	0.68
Particulate matter air pollution 2 5 10um 2010	0.67
Irritability	0.53
Inverse distance to the nearest road	0.48
Mg concentration	0.45
Frequency of tiredness lethargy in last 2 weeks	0.44
Number of full sisters	0.34
Father still alive	0.31
Number of full brothers	0.30
Total traffic load on major roads	0.29
Other serious medical condition disability diagnosed by doctor	0.29
Mother still alive	0.26
Sum of road length of major roads within 100m	0.26
Frequency of depressed mood in last 2 weeks	0.23
Frequency of tenseness restlessness in last 2 weeks	0.22
Nervous feelings	0.20
Loneliness isolation	0.20
Risk taking	0.20
Guilty feelings	0.19
Snoring	0.18
Hearing difficulty problems	0.18
Comparative height size at age 10	0.17
Weight change compared with 1 year ago	0.16
Comparative body size at age 10	0.16
Worry too long after embarrassment	0.15
Illnessmother-heartdisease	0.14

Tense highly strung	0.14
Illnessfather-parkinsons	0.14
Wheeze or whistling in the chest in last year	0.14
Illnessfather-heartdisease	0.13
Illnessfather-emphysema	0.12
Worrier anxious feelings	0.11
Illnessmother-highbloodpressure	0.10
Illnessmother-dementia	0.10
Mood swings	0.10
Hearing difficulty problems with background noise	0.09
Illnessmother-emphysema	0.08
Miserableness	0.08
Illnessmother-parkinsons	0.08
Illnessfather-stroke	0.08
Sensitivity hurt feelings	0.08
Chest pain or discomfort	0.07
Illnessfather-lungcancer	0.07
Illnessfather-prostatecancer	0.06
Illnessmother-stroke	0.06
Illnessfather-highbloodpressure	0.06
Illnessmother-severedepression	0.06
Illnessfather-diabetes	0.05
Illnessmother-diabetes	0.04
Illnessfather-bowelcancer	0.04
Illnessmother-breastcancer	0.04
Close to major road	0.01

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from bag 2

Table S2.3 - The most important variables that had NA values 10% < NA% ≤15%

Variable	Importance
Place of birth in UK east co ordinate	5.97
Place of birth in UK north co ordinate	5.77
Age first had sexual intercourse	3.81
Duration of walks	3.72
Average total household income before tax	3.13
Index of Multiple Deprivation England	2.67
Housing score England	2.32
Education score England	2.29
Age started wearing glasses or contact lenses	2.28
Living environment score England	1.81
Health score England	1.66
Domestic garden percentage buffer 300m	1.53
Greenspace percentage buffer 300m	1.49
Illnesssibling-parkinsons	1.41
Water percentage buffer 1000m	1.31
Domestic garden percentage buffer 1000m	1.25

Crime score England	1.24
Illnesssibling-highbloodpressure	0.91
Water percentage buffer 300m	0.90
Morning evening person chronotype	0.84
Diagnoses main ICD10_gonarthrosis	0.82
Diagnoses main ICD10_inguinalhernia	0.67
Maternal smoking around birth	0.51
Employment score England	0.49
Diagnoses main ICD10_cataract	0.33
Diagnoses main ICD10_atherosclerotic_heartdisease	0.18
Diagnoses main ICD10_colonpolyp	0.15
Diagnoses main ICD10_diverticulardiseaselargeintestine	0.13
illnesssibling_heartdisease	0.12
Diagnoses main ICD10_morbidity	0.11
illnesssibling_diabetes	0.09
illnesssibling_severedepression	0.08

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from bag 3

Table S2.4: The most important variables that had NA values 15% < NA% ≤65%

Variable	Importance
Duration of other exercises	4.22
Frequency of other exercises in last 4 weeks	4.22
Length of working week for main job	3.55
Average weekly beer plus cider intake	2.82
Number of children fathered	1.67
Age at first live birth	1.45
Neuroticism score	1.42
Duration of vigorous activity	1.36
household_relations-child	1.10
Lifetime number of sexual partners	1.09
Weekly usage of mobile phone in last 3 months	1.04
Difference in mobile phone use compared to two years previously	1.00
Frequency of travelling from home to job workplace	1.00
household_relations-partner	0.99
Mother's age	0.96
Time employed in main current job	0.83
Average weekly red wine intake	0.79
Had menopause	0.76
Birth weight	0.70
Distance between home and job workplace	0.69
Duration of light DIY	0.67
Job code at visit	0.66
Birth weight of first child	0.64
Frequency of walking for pleasure in last 4 weeks	0.64
Average weekly spirits intake	0.63

Duration of moderate activity	0.62
Mother s age at death	0.62
Father s age at death	0.60
transport_work-motor_vehicle	0.58
Age at last live birth	0.56
Alcohol usually taken with meals	0.54
Summed days activity	0.53
MET minutes per week for vigorous activity	0.53
Summed MET minutes per week for all activity	0.52
Duration walking for pleasure	0.52
Frequency of light DIY in last 4 weeks	0.49
Average weekly champagne plus white wine intake	0.48
MET minutes per week for moderate activity	0.48
Number of live births	0.47
Age when periods started menarche	0.46
Age completed full time education	0.45
Ever had breast cancer screening mammogram	0.40
Duration of heavy DIY	0.39
MET minutes per week for walking	0.37
Summed minutes activity	0.36
Usual side of head for mobile phone use	0.35
Age started oral contraceptive pill	0.32
Job involves shift work	0.32
Hands free device speakerphone use with mobile phone in last 3 month	0.31
Job involves heavy manual or physical work	0.30
Job involves mainly walking or standing	0.29
Average weekly fortified wine intake	0.27
Age when last used oral contraceptive pill	0.27
Years since last cervical smear test	0.22
transport_work-cycle	0.22
Breastfed as a baby	0.21
Reason for reducing amount of alcohol drunk	0.19
Frequency of heavy DIY in last 4 weeks	0.17
transport_work-public_transport	0.13
Hair balding pattern	0.13
Ever had prostate specific antigen PSA test	0.12
Ever had stillbirth spontaneous miscarriage or termination	0.11
Ever used hormone replacement therapy HRT	0.10
Relative age voice broke	0.09
Medication-bloodpressure	0.09
transport_work-walk	0.09
Medication-cholesterol	0.08
Relative age of first facial hair	0.08
Had major operations	0.08
Bilateral oophorectomy both ovaries removed	0.07
Ever taken oral contraceptive pill	0.07
At or above moderate vigorous recommendation	0.06

Had other major operations	0.06
Medication-cholesterol	0.06
IPAQ activity group	0.06
Medication-HRT	0.05
Medication-bloodpressure	0.05
Ever had hysterectomy womb removed	0.05

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from bag 4

Table S2.5: The most important variables that had NA values 65% < NA% ≤70%

Variable	**Importance**
Friendships satisfaction	8.69
Age at menopause last menstrual period	7.95
Years since last breast cancer screening mammogram	5.20
Belief that own life is meaningful	2.41
Work job satisfaction	1.99
Scale to indicate how health is today	1.48
Been in a confiding relationship as an adult	1.29
Happiness	1.28
Family relationship satisfaction	1.27
Financial situation satisfaction	1.14
Health satisfaction	1.12
General happiness	0.95
Felt loved as a child	0.90
Number of older siblings	0.74
Ever taken cannabis	0.70
Private healthcare	0.57
Eye-cataract	0.47
General happiness with own health	0.42
Heating-gas	0.42
Shortness of breath walking on level ground	0.38
Felt very upset when reminded of stressful experience in past month	0.37
Mobility problems today	0.36
Ever contemplated self harm	0.35
Pain discomfort today	0.34
Anxiety depression today	0.31
Noisy workplace	0.31
Someone to take to doctor when needed as a child	0.30
Belittlement by partner or ex partner as an adult	0.28
Physically abused by family as a child	0.28
Ever depressed for a whole week	0.27
Avoided activities or situations because of previous stressful experience in past month	0.26
Diagnosed with life threatening illness	0.26
Self care problems today	0.25

Tinnitus	0.24
Recent feelings of tiredness or low energy	0.23
Physical violence by partner or ex partner as an adult	0.23
Able to pay rent mortgage as an adult	0.23
Felt hated by family member as a child	0.23
Recent trouble concentrating on things	0.23
Loud music exposure frequency	0.22
Recent worrying too much about different things	0.21
Ever manic hyper for 2 days	0.21
Recent easy annoyance or irritability	0.21
Recent lack of interest or pleasure in doing things	0.21
Recent feelings or nervousness or anxiety	0.21
Ever unenthusiastic disinterested for a whole week	0.20
Recent inability to stop or control worrying	0.19
Trouble falling or staying asleep or sleeping too much	0.19
Recent feelings of depression	0.19
Repeated disturbing thoughts of stressful experience in past month	0.19
Recent feelings of inadequacy	0.18
Leg pain on walking	0.18
Recent poor appetite or overeating	0.18
Recent restlessness	0.18
Non accidental death in close genetic family	0.18
Ever had prolonged feelings of sadness or depression	0.17
Ever highly irritable argumentative for 2 days	0.17
Recent trouble relaxing	0.16
Recent feelings of foreboding	0.16
Ever thought that life not worth living	0.16
Eye-other	0.16
Ever sought or received professional help for mental distress	0.16
Ever had period extreme irritability	0.13
Witnessed sudden violent death	0.13
Sexually molested as a child	0.12
Victim of sexual assault	0.12
Victim of physically violent crime	0.12
Been in serious accident believed to be life threatening	0.11
Ever addicted to any substance or behaviour	0.11
Ever suffered mental distress preventing usual activities	0.10
Ever had prolonged loss of interest in normal activities	0.09

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from bag 5

Table S2.6 Final 75 predictors of Social Participation (XGBoost Importance Ranking)

Variable	Bag ID	Importance

Duration of other exercises	4	3.24
Frequency of other exercises in last 4 weeks	4	2.98
Length of working week for main job	4	2.60
Average weekly beer plus cider intake	4	2.47
Able to confide	2	2.31
Age at first live birth	4	1.95
Employment-retired	1	1.79
Number of children fathered	4	1.68
Place of birth in UK east co ordinate	3	1.27
Friendships satisfaction	5	1.18
Number in household	1	1.09
Place of birth in UK north co ordinate	3	1.07
Number of days week of vigorous physical activity 10 minutes	2	0.90
Frequency of travelling from home to job workplace	4	0.86
Mother s age	4	0.83
transport_leisure-motor	1	0.77
Weekly usage of mobile phone in last 3 months	4	0.76
Length of time at current address	1	0.71
Activity-otherexercises	1	0.71
Time spent watching television TV	1	0.68
household_relations-partner	4	0.66
Difference in mobile phone use compared to two years previously	4	0.65
Distance Euclidean to coast	1	0.61
Time spent driving	2	0.61
Activity-strenuous sports	1	0.60
Age at recruitment	1	0.58
Ca concentration	2	0.57
Length of mobile phone use	2	0.57
Number of days week of moderate physical activity 10 minutes	2	0.56
Duration of vigorous activity	4	0.54
Lifetime number of sexual partners	4	0.54
Fed up feelings	2	0.53
Health score England	3	0.50
Townsend deprivation index at recruitment	1	0.50
Housing score England	3	0.49
Number of days week walked 10 minutes	2	0.49
Nitrogen dioxide air pollution 2010	2	0.48
Nitrogen oxides air pollution 2010	2	0.47
Average 24-hour sound level of noise pollution	2	0.44
Traffic intensity on the nearest major road	2	0.43
Living environment score England	3	0.43
Natural environment percentage buffer 1000m	1	0.43
Genetic sex	2	0.42
Alcohol intake frequency	1	0.42
household_relations-child	4	0.41

Education score England	3	0.40
Water percentage buffer 1000m	3	0.40
Domestic garden percentage buffer 300m	3	0.40
Index of Multiple Deprivation England	3	0.39
Activity-walking	1	0.39
Age first had sexual intercourse	3	0.38
Age started wearing glasses or contact lenses	3	0.36
Greenspace percentage buffer 300m	3	0.36
Belief that own life is meaningful	5	0.34
Duration of walks	3	0.34
Age at menopause last menstrual period	5	0.32
Neuroticism score	4	0.32
Time spent using computer	1	0.27
Scale to indicate how health is today	5	0.26
Own or rent accommodation lived in	2	0.25
Average total household income before tax	3	0.22
Frequency of unenthusiasm disinterest in last 2 weeks	2	0.21
Employment-employed	1	0.20
Felt loved as a child	5	0.17
Years since last breast cancer screening mammogram	5	0.16
Been in a confiding relationship as an adult	5	0.16
Family relationship satisfaction	5	0.15
Illnesssibling-parkinsons	3	0.14
Number of older siblings	5	0.10
Happiness	5	0.10
Ever taken cannabis	5	0.09
Health satisfaction	5	0.08
Work job satisfaction	5	0.08
Financial situation satisfaction	5	0.08
General happiness	5	0.08

Legend: The importance value of the variables in an XGBoost analysis where the input variables were selected from the most important 15 variables from each of the previous bags combined. While Bag 5 variables ranked lower overall, several from Bag 4 scored highly, indicating that predictor importance reflected genuine associations rather than missingness.

Appendix 3

Table S3.1: Regression Results

Predictor	Adjusted R^2	Adjusted R^2 with Age/Sex	Coef.	P-value	Number of cases	Number of cases with Age/Sex
Activity-other exercises*	0.0473	0.0825	0.2290	0	393021	384765
Friendships satisfaction	0.0450	0.0670	-0.1930	0	134667	130467
Belief that own life is meaningful	0.0166	0.0510	0.1210	0	121490	119424
Duration of other exercises*	0.0040	0.0508	0.0607	4.38E-162	186116	182572
Frequency of other exercises in last 4 weeks*	0.0026	0.0501	0.0541	3.75E-130	185921	182381
Number of days week of vigorous physical activity 10 minutes*	0.0107	0.0474	0.1240	0	374221	366455
Able to confide	0.0176	0.0472	0.1290	0	381036	373091
Activity-strenuous sports*	0.0080	0.0467	0.1290	0	393021	384765
Happiness	0.0191	0.0464	-0.1290	0	135709	131468
Family relationship satisfaction	0.0183	0.0459	-0.1250	0	135013	130806
General happiness	0.0120	0.0449	-0.0954	1.68E-250	123753	121656
Employment-retired	0.0290	0.0430	0.1500	0	394300	385985
Number of days week of moderate physical activity 10 minutes*	0.0134	0.0428	0.1050	0	374502	366728
Felt loved as a child	0.0071	0.0427	0.0820	9.34E-189	123972	121873
Been in a confiding relationship as an adult	0.0053	0.0416	0.0739	4.66E-150	121219	119166
Alcohol intake frequency	0.0065	0.0403	-0.1010	0	393997	385697
Fed up feelings	0.0102	0.0393	-0.0936	0	385616	377533
Transport_leisure-motor	0.0079	0.0393	0.0931	0	392825	384574
Frequency of unenthusiasm disinterest in last 2 weeks	0.0109	0.0390	-0.0909	0	380205	372290
Duration of vigorous activity*	0.0000	0.0389	0.0177	5.52e-17	214314	210205
Number of days week walked 10 minutes*	0.0103	0.0389	0.0890	0	387206	379117
Financial situation satisfaction	0.0137	0.0388	-0.0944	4.86E-253	135403	131183
Neuroticism score	0.0064	0.0384	-0.0885	0	316932	310423
Health satisfaction	0.0084	0.0379	-0.0882	8.18E-229	135726	131490
Length of mobile phone use	0.0017	0.0376	0.0846	0	389261	381097
Activity-walking*	0.0085	0.0374	0.0825	0	393021	384765
Employment-employed	0.0203	0.0372	-0.0974	0	394300	385985
Weekly usage of mobile phone in last 3 months	0.0000	0.0365	0.0467	2.11E-151	330909	323917

Difference in mobile phone use compared to two years previously	0.0012	0.0362	0.0440	1.42E-144	332308	325267
Ever taken cannabis	0.0039	0.0361	-0.0184	2.29e-10	124290	122179
Own or rent accommodation lived in	0.0131	0.0355	-0.0746	0	388696	380565
Scale to indicate how health is today	0.0002	0.0350	0.0204	7.47e-14	131768	129523
Average weekly beer plus cider intake	0.0001	0.0348	0.0848	0	272336	267092
Time spent watching television TV	0.0018	0.0348	-0.0649	0	391331	383115
Age started wearing glasses or contact lenses	0.0008	0.0343	0.0144	1.29e-16	337380	330335
Lifetime number of sexual partners	0.0000	0.0340	0.0050	4.10e-3	319722	313364
Townsend deprivation index at recruitment	0.0048	0.0337	-0.0562	5.44E-272	392679	384402
Ca concentration	0.0037	0.0335	-0.0586	3.44E-271	357471	350097
Housing score England	0.0040	0.0335	-0.0592	1.75E-263	338955	331828
Illness_sibling-parkinsons	0.0023	0.0329	-0.0465	1.41E-164	341337	334161
Number in household	0.0074	0.0324	-0.0445	7.70E-152	391741	383513
Length of time at current address	0.0083	0.0323	0.0450	1.33E-146	393104	384825
Place of birth in UK north co ordinate	0.0022	0.0322	0.0437	5.51E-152	350828	344005
Living environment score England	0.0033	0.0321	-0.0457	1.95E-156	338955	331828
Distance Euclidean to coast	0.0012	0.0320	-0.0355	1.59E-110	390384	382177
Age first had sexual intercourse	0.0002	0.0320	-0.0087	4.18e-7	343647	336757
Number of older siblings	0.0017	0.0318	-0.0376	2.59e-38	119342	115613
Place of birth in UK east co ordinate	0.0014	0.0316	-0.0359	2.73E-103	350828	344005
Time spent using computer	0.0031	0.0316	-0.0294	1.21e-74	391110	382887
Time spent driving	0.0002	0.0315	0.0286	1.26e-67	387951	379825
Average total household income before tax	0.0036	0.0315	-0.0118	8.61e-11	338443	331562
Duration of walks*	0.0000	0.0314	0.0059	5.81e-4	338077	331145
Natural environment percentage buffer 1000m	0.0010	0.0313	0.0219	9.18e-43	390384	382177
Nitrogen dioxide air pollution 2010	0.0011	0.0310	-0.0209	3.37e-39	388470	380324
Index of Multiple Deprivation England	0.0020	0.0310	-0.0323	3.46e-79	338955	331828
Nitrogen oxides air pollution 2010	0.0007	0.0309	-0.0170	1.57e-26	388470	380324
Average 24-hour sound level of noise pollution	0.0002	0.0307	-0.0095	2.19e-9	388470	380324

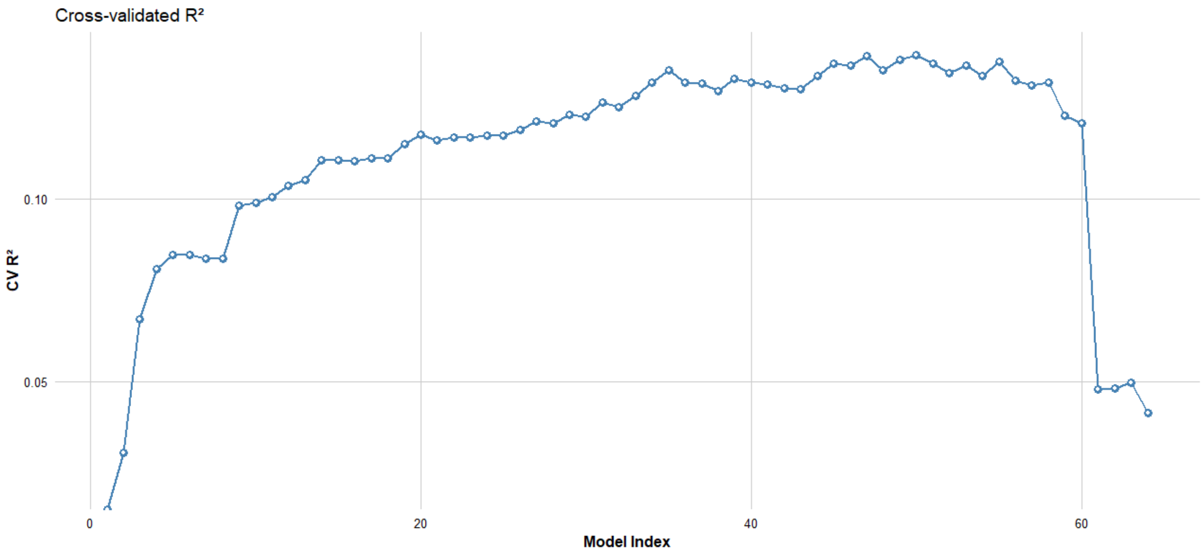
Traffic intensity on the nearest major road	0.0001	0.0306	-0.0068	1.85e-5	388470	380324
Greenspace percentage buffer 300m	0.0005	0.0305	0.0153	1.74e-19	345420	338196
Education score England	0.0009	0.0304	-0.0209	2.58e-34	338955	331828
Domestic garden percentage buffer 300m	0.0002	0.0304	0.0122	5.61e-13	345420	338196
Work job satisfaction	0.0029	0.0303	0.0081	5.15e-3	135602	131365
Water percentage buffer 1000m	0.0000	0.0303	-0.0065	1.39e-4	345420	338196
household_relations-child	0.0079	0.0303	-0.0371	1.85e-78	320256	313752
Health score England	0.0000	0.0301	0.0136	1.87e-15	338043	330932
household_relations-partner	0.0001	0.0292	0.0029	1.03e-1	320256	313752
Mother s age	0.0136	0.0282	0.0439	1.54e-26	155513	152278
Age at first live birth	0.0014	0.0246	-0.0103	6.34e-5	145592	142171
Age at menopause last menstrual period	0.0015	0.0203	0.0152	1.04e-7	121990	118997
Years since last breast cancer screening mammogram	0.0026	0.0198	-0.0195	6.78e-12	126582	123567
Length of working week for main job	0.0143	0.0187	-0.0900	0	222646	218171
Frequency of travelling from home to job workplace	0.0039	0.0140	-0.0496	6.66E-125	221648	217183
Number of children fathered	0.0051	0.0135	0.0578	3.32E-127	177927	174952

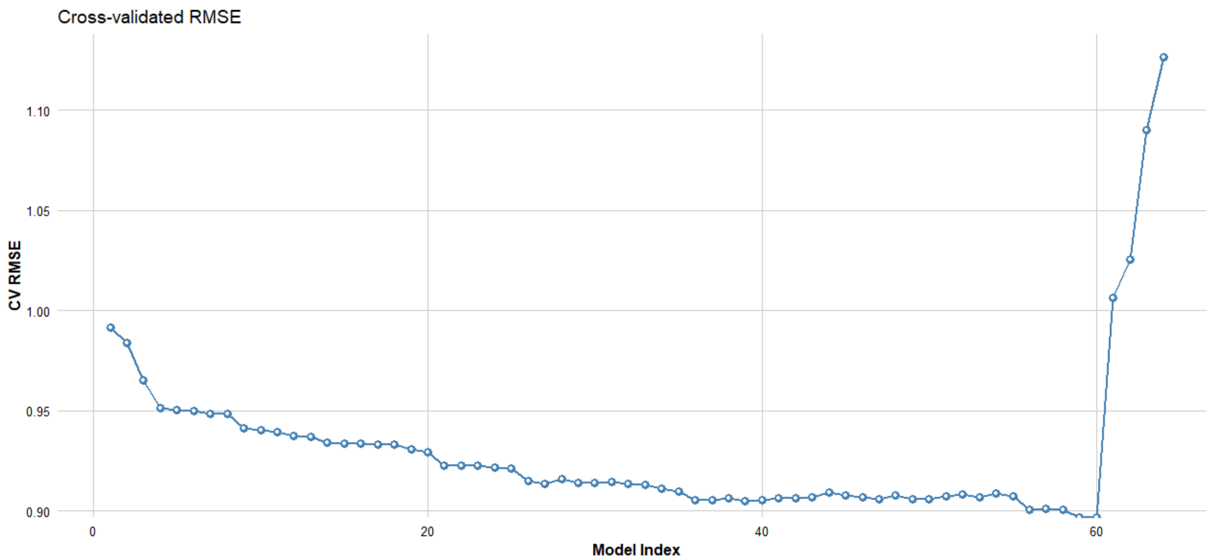
Legend: Regression results showing R^2 of individual regressions and when combined with age/sex; also showing effect size, direction of effect, p-value and number of cases for individual regression and when combined with age/sex. * indicates auto-correlate

Appendices:

Regression Tables and Plots

Full Regression model Adj R^2 and RMSE plots





Appendix 4

Table S4.1: Wellbeing factors in top 75

Able to confide
Been in a confiding relationship as an adult
Belief that own life is meaningful
Family relationship satisfaction
Fed up feelings
Felt loved as a child
Financial situation satisfaction
Frequency of unenthusiasm disinterest in last 2 weeks
Friendships satisfaction 0
General happiness
Happiness
Health satisfaction
Neuroticism score
Work job satisfaction

Appendix 5

Table S5.1: Moderation Analysis Detailed Results

Full predictive model:

```
****************** PART 1. Regression Model Summary ***************
***

PROCESS Model Code : 1 (Hayes, 2018; www.guilford.com/p/hayes3)
PROCESS Model Type : Simple Moderation
-     Outcome (Y) : Edited_Social
-   Predictor (X) : full
-   Mediators (M) : -
- Moderators (W) : Psychosis_Status
- Covariates (C) : Age.at.recruitment, Genetic.sex
-    HLM Clusters : -

All numeric predictors have been grand-mean centered.
(For details, please see the help page of PROCESS.)

Formula of Outcome:
-    Edited_Social ~ Age.at.recruitment + Genetic.sex + full*Psychos
is_Status

CAUTION:
  Fixed effect (coef.) of a predictor involved in an interaction
  denotes its "simple effect/slope" at the other predictor = 0.
  Only when all predictors in an interaction are mean-centered
  can the fixed effect denote the "main effect"!

Model Summary
```

	(1) Edited_Social	(2) Edited_Social
(Intercept)	4.228 ***	4.227 ***
	(0.007)	(0.007)
Age.at.recruitment	0.000	0.000
	(0.001)	(0.001)
Genetic.sex	-0.000	0.000
	(0.015)	(0.015)
full	1.000 ***	1.000 ***
	(0.016)	(0.016)
Psychosis_Status1		0.137
		(0.093)
full:Psychosis_Status1		0.139
		(0.151)
R^2	0.115	0.115
Adj. R^2	0.115	0.115
Num. obs.	42564	42564

Note. * *p* < .05, ** *p* < .01, *** *p* < .001.

```
************ PART 2. Mediation/Moderation Effect Estimate *********
***

Package Use : ‘interactions’ (v1.2.0)
Effect Type : Simple Moderation (Model 1)
Sample Size : 42564 (351736 missing observations deleted)
Random Seed : -
Simulations : -

Interaction Effect on "Edited_Social" (Y)
```

	F	df1	df2	p

```
full * Psychosis_Status  0.85   1 42558  .358
```

```
Simple Slopes: "full" (X) ==> "Edited_Social" (Y)
```

"Psychosis_Status"	Effect	S.E.	t	p		[95% CI]
0	1.000	(0.016)	60.771	<.001	***	[0.967, 1.032]
1	1.139	(0.151)	7.548	<.001	***	[0.843, 1.434]

Parsimonious model

```
****************** PART 1. Regression Model Summary ***************
***

PROCESS Model Code : 1 (Hayes, 2018; www.guilford.com/p/hayes3)
PROCESS Model Type : Simple Moderation
-    Outcome (Y) : Edited_Social
-  Predictor (X) : parsimonious
-  Mediators (M) : -
- Moderators (W) : Psychosis_Status
- Covariates (C) : Age.at.recruitment, Genetic.sex
-   HLM Clusters : -

All numeric predictors have been grand-mean centered.
(For details, please see the help page of PROCESS.)

Formula of Outcome:
-    Edited_Social ~ Age.at.recruitment + Genetic.sex + parsimonious
*Psychosis_Status

CAUTION:
  Fixed effect (coef.) of a predictor involved in an interaction
  denotes its "simple effect/slope" at the other predictor = 0.
  Only when all predictors in an interaction are mean-centered
  can the fixed effect denote the "main effect"!

Model Summary
```

	(1) Edited_Social	(2) Edited_Social
(Intercept)	4.225 *** (0.007)	4.225 *** (0.007)
Age.at.recruitment	0.000 (0.001)	-0.000 (0.001)
Genetic.sex	-0.000 (0.015)	0.000 (0.015)
parsimonious	1.000 *** (0.018)	1.001 *** (0.018)
Psychosis_Status1		0.110 (0.093)
parsimonious:Psychosis_Status1		0.031 (0.150)
R^2	0.102	0.102
Adj. R^2	0.102	0.102
Num. obs.	43832	43832

Note. * p < .05, ** p < .01, *** p < .001.

```
************ PART 2. Mediation/Moderation Effect Estimate *********
***

Package Use : 'interactions' (v1.2.0)
Effect Type : Simple Moderation (Model 1)
```

```
Sample Size : 43832 (350468 missing observations deleted)
Random Seed : -
Simulations : -

Interaction Effect on "Edited_Social" (Y)
```

	F	df1	df2	p
parsimonious * Psychosis_Status	0.04	1	43826	.839

Simple Slopes: "parsimonious" (X) ==> "Edited_Social" (Y)

"Psychosis_Status"	Effect	S.E.	t	p		[95% CI]
0	1.001	(0.018)	55.725	<.001	***	[0.965, 1.036]
1	1.031	(0.150)	6.878	<.001	***	[0.737, 1.325]

Latent Factor:

```
***************** PART 1. Regression Model Summary ****************
**

PROCESS Model Code : 1 (Hayes, 2018; www.guilford.com/p/hayes3)
PROCESS Model Type : Simple Moderation
-    Outcome (Y) : Edited_Social
-  Predictor (X) : latent_factor
-  Mediators (M) : -
- Moderators (W) : Psychosis_Status
- Covariates (C) : Age.at.recruitment, Genetic.sex
-   HLM Clusters : -

All numeric predictors have been grand-mean centered.
(For details, please see the help page of PROCESS.)

Formula of Outcome:
-    Edited_Social ~ Age.at.recruitment + Genetic.sex + latent_facto
r*Psychosis_Status

CAUTION:
  Fixed effect (coef.) of a predictor involved in an interaction
  denotes its "simple effect/slope" at the other predictor = 0.
  Only when all predictors in an interaction are mean-centered
  can the fixed effect denote the "main effect"!

Model Summary
```

	(1) Edited_Social	(2) Edited_Social
(Intercept)	4.224 ***	4.223 ***
	(0.007)	(0.007)
Age.at.recruitment	0.025 ***	0.025 ***
	(0.001)	(0.001)
Genetic.sex	-0.328 ***	-0.328 ***
	(0.014)	(0.014)
latent_factor	1.085 ***	1.083 ***
	(0.023)	(0.023)
Psychosis_Status1		0.124
		(0.094)
latent_factor:Psychosis_Status1		0.376

		(0.244)
R^2	0.084	0.084
Adj. R^2	0.083	0.083
Num. obs.	43079	43079

Note. * *p* < .05, ** *p* < .01, *** *p* < .001.

**************** PART 2. Mediation/Moderation Effect Estimate ************

Package Use : ‘interactions’ (v1.2.0)
Effect Type : Simple Moderation (Model 1)
Sample Size : 43079 (351221 missing observations deleted)
Random Seed : -
Simulations : -

Interaction Effect on "Edited_Social" (Y)

	F	df1	df2	p
latent_factor * Psychosis_Status	2.38	1	43073	.123

Simple Slopes: "latent_factor" (X) ==> "Edited_Social" (Y)

"Psychosis_Status"	Effect	S.E.	t	p		[95% CI]
0	1.083	(0.023)	46.393	<.001	***	[1.037, 1.128]
1	1.459	(0.243)	6.008	<.001	***	[0.983, 1.934]

References

Allen, S. F., Gilbody, S., Atkin, K., & van der Feltz-Cornelis, C. M. (2023). The associations among childhood trauma, loneliness, mental health symptoms, and indicators of social exclusion in adulthood: A UK Biobank study. *Brain and Behavior*, *13*(4), e2959. https://doi.org/10.1002/brb3.2959

Ashida, T., Kondo, N., & Kondo, K. (2016). Social participation and the onset of functional disability by socioeconomic status and activity type: The JAGES cohort study. *Preventive Medicine*, *89*, 121–128. https://doi.org/10.1016/j.ypmed.2016.05.006

Brown, T. A. (2015). *Confirmatory factor analysis for applied research, 2nd ed* (pp. xvii, 462). The Guilford Press.

Cacciatore, S., Mao, S., Nuñez, M. V., Massaro, C., Spadafora, L., Bernardi, M., Perone, F., Sabouret, P., Biondi-Zoccai, G., Banach, M., Calvani, R., Tosato, M., Marzetti, E., & Landi, F. (2025). Urban health inequities and healthy longevity: Traditional and emerging risk factors across the cities and policy implications. *Aging Clinical and Experimental Research*, *37*(1), 143. https://doi.org/10.1007/s40520-025-03052-1

Cao, S., Liu, L., Zhu, Q., Zhu, Z., Zhou, J., Wei, P., & Wu, M. (2021). Association Between Dietary Patterns and Plasma Lipid Biomarker and Female Breast Cancer Risk: Comparison of Latent Class Analysis (LCA) and Factor Analysis (FA). *Frontiers in Nutrition*, *8*. https://doi.org/10.3389/fnut.2021.645398

Casburn, M., Laighneach, A., Doherty, E., & Quilligan, F. (2025). The Brain Network Underlying Social Participation: A Multimodal, Data-Driven Investigation. *Brain Imaging and Behaviour, submitted*.

Charlotte, S., & Bello, S. (2023). *Evaluating the Performance of Ensemble Methods on Imbalanced Datasets: A Deep Dive into Bagging, Boosting, and Stacking*. 25.

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794. https://doi.org/10.1145/2939672.2939785

Cohen, S., & Wills, T. A. (1985). Stress, social support, and the buffering hypothesis. *Psychological Bulletin*, *98*(2), 310–357. https://doi.org/10.1037/0033-2909.98.2.310

Decade of healthy ageing. (2021). World Health Organization. https://www.who.int/publications/b/56512

Degnan, A., Berry, K., Sweet, D., Abel, K., Crossley, N., & Edge, D. (2018). Social networks and symptomatic and functional outcomes in schizophrenia: A systematic review and meta-analysis. *Social Psychiatry and Psychiatric Epidemiology*, *53*(9), 873–888. https://doi.org/10.1007/s00127-018-1552-8

Doherty, E., Laighneach, A., Casburn, M., Quilligan, F., Donohoe, G., Cannon, D. M., & Morris, D. W. (2025). *Genome-wide association studies of social participation and occupational engagement in the UK Biobank* (p.

2025.05.15.25327674). medRxiv. https://doi.org/10.1101/2025.05.15.25327674

Duangsoithong, R., & Windeatt, T. (2010). Bootstrap feature selection for ensemble classifiers. *Proceedings of the 10th Industrial Conference on Advances in Data Mining: Applications and Theoretical Aspects*, 28–41.

Enders, C. K. (2010). *Applied missing data analysis* (pp. xv, 377). The Guilford Press.

Ergul Aydin, Z., & Kamisli Ozturk, Z. (2021). *Performance Analysis of XGBoost Classifier with Missing Data*.

Flegal, K. M., Kit, B. K., Orpana, H., & Graubard, B. I. (2013). Association of All-Cause Mortality With Overweight and Obesity Using Standard Body Mass Index Categories: A Systematic Review and Meta-analysis. *JAMA*, *309*(1), 71–82. https://doi.org/10.1001/jama.2012.113905

Flores-Kanter, P. E., Garrido, L. E., Moretti, L. S., & Medrano, L. A. (2021). A modern network approach to revisiting the Positive and Negative Affective Schedule (PANAS) construct validity. *Journal of Clinical Psychology*, *77*(10), 2370–2404. https://doi.org/10.1002/jclp.23191

Forster, M. R. (2000). Key Concepts in Model Selection: Performance and Generalizability. *Journal of Mathematical Psychology*, *44*(1), 205–231. https://doi.org/10.1006/jmps.1999.1284

Foster, H. M. E., Gill, J. M. R., Mair, F. S., Celis-Morales, C. A., Jani, B. D., Nicholl, B. I., Lee, D., & O'Donnell, C. A. (2023). Social connection and mortality in UK Biobank: A prospective cohort analysis. *BMC Medicine*, *21*(1), 384. https://doi.org/10.1186/s12916-023-03055-7

Fratiglioni, L., Paillard-Borg, S., & Winblad, B. (2004). An active and socially integrated lifestyle in late life might protect against dementia. *The Lancet. Neurology*, *3*(6), 343–353. https://doi.org/10.1016/S1474-4422(04)00767-7

Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, *29*(5), 1189–1232. https://doi.org/10.1214/aos/1013203451

Fry-Nartey, L. N., Akafia, C., Nkonu, U. S., Baiden, S. B., Dorvi, I. N., Agyenkwa-Mawuli, K., Agyapong, O., Hayford, C. F., Wilson, M. D., Miller, W. A., & Kwofie, S. K. (2025). AICpred: Machine Learning-Based Prediction of Potential Anti-Inflammatory Compounds Targeting TLR4-MyD88 Binding Mechanism. *Information*, *16*(1), 34. https://doi.org/10.3390/info16010034

Gagiu, C., Dionisie, V., Manea, M. C., Mazilu, D. C., & Manea, M. (2024). Internalised Stigma, Self-Esteem and Perceived Social Support as Psychosocial Predictors of Quality of Life in Adult Patients with Schizophrenia. *Journal of Clinical Medicine*, *13*(22), 6959. https://doi.org/10.3390/jcm13226959

Harrison, G., Hopper, K., Craig, T., Laska, E., Siegel, C., Wanderling, J., Dube, K. C., Ganev, K., Giel, R., Heiden, W. A. D., Holmberg, S. K., Janca, A., Lee, P. W. H., León, C. A., Malhotra, S., Marsella, A. J., Nakane, Y., Sartorius, N., Shen, Y., … Wiersma, D. (2001). Recovery from psychotic illness: A 15- and 25-year international follow-up study. *The British Journal of Psychiatry*, *178*(6), 506–517. https://doi.org/10.1192/bjp.178.6.506

Hastie, T., Friedman, J., & Tibshirani, R. (2001a). Additive Models, Trees, and Related Methods. In T. Hastie, J. Friedman, & R. Tibshirani (Eds.), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (pp. 257–298). Springer. https://doi.org/10.1007/978-0-387-21606-5_9

Hastie, T., Friedman, J., & Tibshirani, R. (2001b). *The Elements of Statistical Learning*. Springer. https://doi.org/10.1007/978-0-387-21606-5

Holt-Lunstad, J., Smith, T. B., Baker, M., Harris, T., & Stephenson, D. (2015). Loneliness and social isolation as risk factors for mortality: A meta-analytic review. *Perspectives on Psychological Science: A Journal of the Association for Psychological Science*, *10*(2), 227–237. https://doi.org/10.1177/1745691614568352

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal*, *6*(1), 1–55. https://doi.org/10.1080/10705519909540118

Hughes, K., Bellis, M. A., Hardcastle, K. A., Sethi, D., Butchart, A., Mikton, C., Jones, L., & Dunne, M. P. (2017). The effect of multiple adverse childhood experiences on health: A systematic review and meta-analysis. *The Lancet Public Health*, *2*(8), e356–e366. https://doi.org/10.1016/S2468-2667(17)30118-4

Husain, W., Trabelsi, K., Ammar, A., Saif, Z., & Jahrami, H. (2025). The development and validation of a one-off scale to measure procrastination and precrastination traits in young adults. *BMC Psychology*, *13*, 773. https://doi.org/10.1186/s40359-025-03072-6

Jacobs, D. R., Jr, Adachi, H., Mulder, I., Kromhout, D., Menotti, A., Nissinen, A., Blackburn, H., & for the Seven Countries Study Group. (1999). Cigarette Smoking and Mortality Risk: Twenty-five–Year Follow-up of the Seven Countries Study. *Archives of Internal Medicine*, *159*(7), 733–740. https://doi.org/10.1001/archinte.159.7.733

Jenkins, L., McNeal, T., Drayer, J., & Wang, Q. (2020). Childhood Trauma History and Negative Social Experiences in College. *Journal of Child & Adolescent Trauma*, *14*(1), 103–113. https://doi.org/10.1007/s40653-020-00315-z

Kobrinsky, V., & Siedlecki, K. L. (2022). Mediators of the Relationship Between Adverse Childhood Experiences (ACEs) and Symptoms of Anxiety, Depression, and Suicidality among Adults. *Journal of Child & Adolescent Trauma*, *16*(2), 233–246. https://doi.org/10.1007/s40653-022-00510-0

Ku, B. S., Compton, M. T., Walker, E. F., & Druss, B. G. (2021). Social Fragmentation and Schizophrenia: A Systematic Review. *The Journal of Clinical Psychiatry*, *83*(1), 21r13941. https://doi.org/10.4088/JCP.21r13941

Kuhn, M., & Vaughan, D. (2025). *parsnip: A Common API to Modeling and Analysis Functions*. https://github.com/tidymodels/parsnip

Kuhn, M., & Wickham, H. (2020). *Tidymodels: A collection of packages for modeling and machine learning using tidyverse principles.* https://www.tidymodels.org

Kung, C. S. J., Pudney, S. E., & Shields, M. A. (2022). Economic gradients in loneliness, social isolation and social support: Evidence from the UK

Biobank. *Social Science & Medicine*, *306*, 115122. https://doi.org/10.1016/j.socscimed.2022.115122

Lehning, A., Kim, K., Smith, R., & Choi, M. (2018). Does economic vulnerability moderate the association between transportation mode and social activity restrictions in later life? *Ageing & Society*, *38*(10), 2041–2060. https://doi.org/10.1017/S0144686X17000411

Levasseur, M., Cohen, A. A., Dubois, M.-F., Généreux, M., Richard, L., Therrien, F.-H., & Payette, H. (2015). Environmental Factors Associated With Social Participation of Older Adults Living in Metropolitan, Urban, and Rural Areas: The NuAge Study. *American Journal of Public Health*, *105*(8), 1718–1725. https://doi.org/10.2105/AJPH.2014.302415

Lieberman, J. A., & First, M. B. (2018). Psychotic Disorders. *The New England Journal of Medicine*, *379*(3), 270–280. https://doi.org/10.1056/NEJMra1801490

Linde, J., Schmid, M. T., Ruud, T., Skar-Fröding, R., & Biringer, E. (2023). Social Factors and Recovery: A Longitudinal Study of Patients with Psychosis in Mental Health Services. *Community Mental Health Journal*, *59*(2), 294–305. https://doi.org/10.1007/s10597-022-01007-8

Lipton, Z. C. (2018). The mythos of model interpretability. *Commun. ACM*, *61*(10), 36–43. https://doi.org/10.1145/3233231

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees.

Nature Machine Intelligence, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Mutz, J., Roscoe, C. J., & Lewis, C. M. (2021). Exploring health in the UK Biobank: Associations with sociodemographic characteristics, psychosocial factors, lifestyle and environmental exposures. *BMC Medicine*, *19*(1), 240. https://doi.org/10.1186/s12916-021-02097-z

O'Connell, K., Marsh, A. A., & Seydell-Greenwald, A. (2024). Right hemisphere stroke is linked to reduced social connectedness in the UK Biobank cohort. *Scientific Reports*, *14*, 27293. https://doi.org/10.1038/s41598-024-78351-0

Okbay, A., Wu, Y., Wang, N., Jayashankar, H., Bennett, M., Nehzati, S. M., Sidorenko, J., Kweon, H., Goldman, G., Gjorgjieva, T., Jiang, Y., Hicks, B., Tian, C., Hinds, D. A., Ahlskog, R., Magnusson, P. K. E., Oskarsson, S., Hayward, C., Campbell, A., … Young, A. I. (2022). Polygenic prediction of educational attainment within and between families from genome-wide association analyses in 3 million individuals. *Nature Genetics*, *54*(4), 437–449. https://doi.org/10.1038/s41588-022-01016-z

Phillips, D. L. (1967). Mental Health Status, Social Participation, and Happiness. *Journal of Health and Social Behavior*, *8*(4), 285–291. https://doi.org/10.2307/2948422

Reinhard, M. A., Rek, S. V., Nenov-Matt, T., Barton, B. B., Dewald-Kaufmann, J., Merz, K., Musil, R., Jobst, A., Brakemeier, E.-L., Bertsch, K., & Padberg, F. (2022). Association of loneliness and social network size in adulthood with childhood maltreatment: Analyses of a population-based and a clinical sample. *European Psychiatry: The Journal of the Association of European Psychiatrists*, *65*(1), e55. https://doi.org/10.1192/j.eurpsy.2022.2313

Repke, M. A., & Ipsen, C. (2020). Differences in social connectedness and perceived isolation among rural and urban adults with disabilities. *Disability and Health Journal*, *13*(1), 100829. https://doi.org/10.1016/j.dhjo.2019.100829

Rojas, A. Z., Bakhtiari, S., Aldrich, C., Calo, V. M., & Iannuzzi, M. (2024). XGBoost model for the quantitative assessment of stress corrosion cracking. *Npj Materials Degradation*, *8*(1), 126. https://doi.org/10.1038/s41529-024-00538-7

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, *48*, 1–36. https://doi.org/10.18637/jss.v048.i02

Sathyanarayana, S., & Mohanasundaram, T. (2024). Fit indices in structural equation modeling and confirmatory factor analysis: Reporting guidelines. *Asian Journal of Economics, Business and Accounting*, *24*(7), 561–577.

Shen, C., Rolls, E. T., Cheng, W., Kang, J., Dong, G., Xie, C., Zhao, X.-M., Sahakian, B. J., & Feng, J. (2022). Associations of Social Isolation and Loneliness With Later Dementia. *Neurology*, *99*(2), e164–e175. https://doi.org/10.1212/WNL.0000000000200583

Tansey, W., Veitch, V., Zhang, H., Rabadan, R., & Blei, D. M. (2022). The Holdout Randomization Test for Feature Selection in Black Box Models. *Journal of*

Computational and Graphical Statistics, *31*(1), 151–162. https://doi.org/10.1080/10618600.2021.1923520

Townsend, B. G., Chen, J. T.-H., & Wuthrich, V. M. (2021). Barriers and Facilitators to Social Participation in Older Adults: A Systematic Literature Review. *Clinical Gerontologist*, *44*(4), 359–380. https://doi.org/10.1080/07317115.2020.1863890

Townsend, P. (1987). Deprivation. *Journal of Social Policy*, *16*(2), 125–146. https://doi.org/10.1017/S0047279400020341

Umberson, D., & Montez, J. K. (2010). Social Relationships and Health: A Flashpoint for Health Policy. *Journal of Health and Social Behavior*, *51*(Suppl), S54–S66. https://doi.org/10.1177/0022146510383501

Vassos, E., Pedersen, C. B., Murray, R. M., Collier, D. A., & Lewis, C. M. (2012). Meta-analysis of the association of urbanicity with schizophrenia. *Schizophrenia Bulletin*, *38*(6), 1118–1123. https://doi.org/10.1093/schbul/sbs096

Wanchai, A., & Phrompayak, D. (2019). A Systematic Review of Factors Influencing Social Participation of Older Adults. *Pacific Rim International Journal of Nursing Research*, *23*(2), 131–141.

Wickramaratne, P. J., Yangchen, T., Lepow, L., Patra, B. G., Glicksburg, B., Talati, A., Adekkanattu, P., Ryu, E., Biernacka, J. M., Charney, A., Mann, J. J., Pathak, J., Olfson, M., & Weissman, M. M. (2022). Social connectedness as a determinant of mental health: A scoping review. *PLOS ONE*, *17*(10), e0275004. https://doi.org/10.1371/journal.pone.0275004

Wilding, A., Munford, L. A., & Sutton, M. (2023). Predictors of social participation: Evidence from repeated cross-sectional population surveys in England. *Journal of Public Health (Oxford, England)*, *45*(2), 379–388. https://doi.org/10.1093/pubmed/fdac029

Wu, P., Ding, J., Jin, H., Zhang, J., Greenshaw, A. J., & Liang, S. (2025). Social activity as a mediator between childhood adversity and depressive symptoms in middle-aged and older Chinese adults. *Frontiers in Psychiatry*, *16*, 1553895. https://doi.org/10.3389/fpsyt.2025.1553895

Yarkoni, T., & Westfall, J. (2017). *Choosing Prediction Over Explanation in Psychology: Lessons From Machine Learning*. https://journals.sagepub.com/doi/10.1177/1745691617693393