**Study Sample and Data Collection**

Our study used a cross-sectional dataset from the Mental Health Million (MHM) project, an on-going study with the purpose of assessing global mental wellbeing through administration of the Mental Health Quotient (MHQ). Initial recruitment targeted the English-speaking population living in the United States, United Kingdom, Canada, South Africa, Singapore, Australia, New Zealand and India, but was later expanded to include Spanish and French speakers as well as other countries for the purpose of capturing a broader global sample. The sample for our present study included 341,956 participants from 229 countries who completed the MHQ from XXX to XXX. Additional information concerning the MHM project and recruitment strategy may be found elsewhere (REF). This study involved secondary analysis of existing data and therefore Institutional Research Ethics Board approval was not required.

**Measures:**

**Mental Wellbeing:**

The MHQ is a 47-item voluntary online survey designed to assess a comprehensive range of common attributes found across widely used existing mental health assessment tools in a single questionnaire to estimate overall mental wellbeing and functioning in the population. The MHQ was developed as a response to the heterogeneity of current psychiatric assessment tools (REF). Items were developed by consolidation of 170 symptoms coded from 126 commonly used assessment tools covering depression, anxiety, bipolar disorder, ADHD, post-traumatic stress disorder, obsessive-compulsive disorder, addiction, schizophrenia, eating disorders and autism spectrum disorder. The MHQ is unique from other psychiatric tools in that the items assessed the level of functioning and impact on one’s life associated with each mental health element, as opposed to frequency, duration, or severity of symptoms. The questionnaire took an average of 14 minutes for participants to complete.

Responses were computed into an overall mental wellbeing score, originally ranging from -100 to +200. Recently, the lower limit was expanded to -166 to accommodate a floor effect. Scores are binned into six levels of functioning, with negative scores indicating clinical risk and positive scores representing normal range: Clinical (≤-50), At Risk (-50 to <0), Enduring (0 to <50), Managing (50 to <100), Succeeding (100 to <150) and Thriving (150 to 200). To compute the overall score, individual item responses were weighted to reflect the nonlinearity of risk associated with increases in symptom severity, as well as the differential risk associated with different symptoms (e.g., suicidal thoughts vs irritability).

In addition to the overall wellbeing score, scores for six broad subcategories of mental wellbeing were computed: Core Cognition (ability for executive functioning), Complex Cognition (reflecting more complex processes such as problem-solving, creativity, and adaptability), Mood and Outlook (ability to effectively regulate ones emotions), Drive and Motivation (ability to achieve goals in the face of obstacles), Social Self (social functioning), and Mind-Body (physical functioning and psychosomatic health). Subcategory scores ranged from -100 to +200, and were computed by a weighted average of scores from 10 to 24 relevant symptom items based on a review of cognitive and brain functioning models (REF).

The MHQ demonstrated high sample reliability when four randomly selected and demographically similar samples were compared on response distributions (p = 0.99), and resulting MHQ distribution (p = 0.18). Internal consistency was demonstrated with conceptually similar items having higher correlations than unsimilar items. A subset of participants which took the MHQ twice at least 3 days apart showed a test-retest reliability of r = 0.84. Validity was assessed by asking a subset of participants additional questions concerning days missed from work and normal activities in the past month. Those who were employed and scored an overall MHQ between 175 to 200 missed on average 0.2 days of work in the past month, while those employed who scored between -75 to -100 missed an average of 9.3 days of work. (REF)

**Physical Activity**: Participants responded to single item that asked: “How regularly do you engage in physical exercise (30 minutes or more)?” Response options included “Rarely/never”; “Less than once a week”; “Once a week”; “Few days a week”; and “Every day.

**Covariates**: To adjust for potential confounders, we selected as covariates age, biological sex, gender identity, ethnicity, educational attainment, employment status, relationship status, frequency of adequate sleep, frequency of socializing, diagnosis of medical condition (Y/N), whether they are currently seeking mental health treatment (Y/N), and whether they reported a significant traumatic childhood or adult experience (Y/N). All items in which participants responded “Prefer not to say” were recoded as missing. Participants were further nested by country in the analysis to account for potential clustering effects.

**Statistical Analysis:**

All data preprocessing and statistical analyses were done using the statistical software R version 4.1.2 (REF). As the items used to calculate the overall MHQ score, as well as exercise frequency, were required by the questionnaire, only some covariate items which were not required showed significant missingness, with ethnicity and gender identity showing the highest (84.2% and 98.5% respectively). These were thus we dropped it from the analysis. Answer responses which included “Prefer not to say” were recoded to missing. In line with Chekroud (REF), responses to physical exercise frequency were recoded into binary groups, with “Rarely/Never” indicating no exposure to exercise, and all other responses indicating exposure to exercise.

Distributions of covariates were balanced between the non-exposure and exposure groups using propensity score weights estimated with generalized boosted modeling (GBM) (REF), and implemented in the R-Package WeightIt (REF). The propensity score is defined as “the conditional probability of assignment to a particular treatment given a vector of observed covariates” (REF Rosen1983) with weighting procedures preferred over matching to preserve the sample size. Propensity scores were converted into weights based on the Average Treatment effect on the Treated estimand, which is used to estimate the hypothetical effect of exercise exposure on the control (i.e. non-exercise) group, and is unbiased when the strong ignorability assumption is met (REF Rosen1983). The GBM is a non-parametric iterative machine learning method which utilizes regression trees and can accommodate non-linearity. As suggested by McCraffrey et al (REF) the number of iterations was determined by minimizing the average standardized absolute mean difference in the covariates. The GBM handles missing data by the surrogate splitting method as described in the WeightIt R-package documentation (REF). Though unlikely to significantly improve performance of inferences for our estimation procedure (REF) weights above 99% were trimmed to reduce potential bias from extreme values. Diagnostics were used to ensure covariates were adequately balanced. Our main analysis estimated treatment effects of exercise on seven outcomes: overall MHQ score, and the six broad subcategories.

In case the inferences of the main analysis are biased due to the method of estimation or handling of missing, we performed a sensitivity analysis with multiple imputation and covariate balancing propensity scores to check for convergence (details in supplementary?). Using this method, we also computed interaction effects between physical exercise and age on the overall MHQ score, with age recoded into three groups: ‘young adult’ (18-34), ‘middle adult’ (35-64), and ‘senior’ (65-85+).

SUPPLEMENTARY

For the sensitivity analysis we handled missing data with Multiple Imputation using Fully Conditioned Specification as implemented in the R-Package *mice* (REF) with number of imputations set to 5. Multiple imputation preserves the sample size while accounting for uncertainty by incorporating randomness in missing value estimation with multiple data sets. Covariate balancing propensity scores (REF) were used to estimate propensity weights on each imputed dataset as implemented in the R-Package *WeightThem* (REF), and the results were combined using the *Within* approach where weights are estimated for each imputed data set, exposure effects are computed for each individual data set and then the coefficients and standard errors are subsequently pooled using Rubin’s Rules (REF) to produce a point estimate of the exposure effect. The within approach demonstrates unbiased estimates when compared to other approaches (Leyrat et al 2019, Granger et al 2019). Further, we also computed a doubly robust estimator by running the regression model with propensity scores and the full covariate set to ensure an unbiased estimate in the case of a misspecified propensity or outcome regression model (REF).