**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 (Newson et al., 2021). 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. Items were developed by consolidation of 170 symptoms coded from 126 commonly used psychiatric 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 (Newson et al., 2020).

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. (Newson JJ, 2022)

**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). ~~These variables will be referred to as the full covariate set~~. 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). 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. For the main analyses,

Distributions of covariates were balanced across exposure levels of exercise frequency using propensity scores computed using generalized boosted models (GBM) (McCaffrey, 2013), and implemented in the R-Package WeightIt (Griefer, 2022). The propensity scores is defined as “the conditional probability of assignment to a particular treatment given a vector of observed covariates” (Rosenbaum & Rubin, 1983). Propensity scores were converted into weights based on the Average Treatment effect on the Treated estimand for binary treatments. The GBM is a non-parametric iterative machine learning method which utilizes regression trees to produce a stable estimation of weights (refs). The GBM accommodates non-linearity and handles missingness by surrogate splitting. Weights above 99% were trimmed to reduce potential bias from extreme weights (ref).

Diagnostics of the covariate distribution balance were used to assess potential model misspecification.. Our main analysis estimated relative? treatment effects of exercise frequency on seven outcomes: overall MHQ score, and the six broad subcategories.

We performed several sensitivity analyses, (the results of which can be found in supplemental?). First, we handled missing data with Multiple Imputation using Fully Conditioned Specification as implemented in the R-Package *mice* (van Buuren 2022) with number of imputations set to 10. Multiple imputation preserves the sample size while accounting for uncertainty by incorporating randomness in missing value estimation with multiple data sets. The GBM was again used to estimate propensity weights on each imputed dataset, and the results were combined using the *Within* approach (Mitra & Reiter, 2016), 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 (Rubin, 1987) 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 (Funk et al, 2011).

We also computed a multiple regression model with the full covariate set without the propensity weights to ensure results were not an artifact of weighting procedures. To adjust for potential collider bias (Holmberg & Anderson, 2022), we then reran the regression model without covariates which could plausibly be influenced by both mental wellbeing and exercise frequency.

Finally, in a secondary analysis we investigated potential interaction effects of exercise frequency and age groups on the overall MHQ score separately in males and females. We split the sample by sex, dropping participants whose responses to “What is your biological sex” corresponded to “Other/Intersex” (n = 819), “Prefer not to say” (n = 2573), or were missing (n = 714) and grouped ages into young adult (18-34), middle adult (35-64), and senior adult (65-85+).

Michele Jonsson Funk, Daniel Westreich, Chris Wiesen, Til Stürmer, M. Alan Brookhart, Marie Davidian, Doubly Robust Estimation of Causal Effects, American Journal of Epidemiology, Volume 173, Issue 7, 1 April 2011, Pages 761–767, https://doi.org/10.1093/aje/kwq439

Holmberg MJ, Andersen LW. Collider Bias. JAMA. 2022;327(13):1282–1283. doi:10.1001/jama.2022.1820

Farhad Pishgar, Noah Greifer, Clémence Leyrat and Elizabeth Stuart (2021). MatchThem:: Matching and Weighting after Multiple Imputation. The R Journal. doi: 10.32614/RJ-2021-073.

Stef van Buuren, Karin Groothuis-Oudshoorn (2011). mice: Multivariate Imputation by Chained Equations in R. Journal of Statistical Software, 45(3), 1-67. DOI 10.18637/jss.v045.i03.

Mitra, R., & Reiter, J. P. (2016). A comparison of two methods of estimating propensity scores after multiple imputation. Statistical methods in medical research, 25(1), 188-204.

Newson, J. J., Pastukh, V., & Thiagarajan, T. C. (2022). Assessment of Population Well-being With the Mental Health Quotient: Validation Study. *JMIR Mental Health*, *9*(4), e34105.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70(1), 41-55.

Leyrat C, Seaman SR, White IR, Douglas I, Smeeth L, Kim J, Resche-Rigon M, Carpenter JR, Williamson EJ. Propensity score analysis with partially observed covariates: How should multiple imputation be used? Stat Methods Med Res. 2019 Jan;28(1):3-19. doi: 10.1177/0962280217713032. Epub 2017 Jun 2. PMID: 28573919; PMCID: PMC6313366.

Rubin, D. B. (2004). Multiple imputation for nonresponse in surveys (Vol. 81). John Wiley & Sons.

Granger, E, Sergeant, JC, Lunt, M. Avoiding pitfalls when combining multiple imputation and propensity scores. Statistics in Medicine. 2019; 38: 5120– 5132. <https://doi.org/10.1002/sim.8355>

McCaffrey, D. F., Ridgeway, G., & Morral, A. R. (2004). Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. Psychological Methods, 9(4), 403–425. doi:10.1037/1082-989X.9.4.403