



## RESEARCH ARTICLE

# Physical activity offsets genetic risk for incident depression assessed via electronic health records in a biobank cohort study

Karmel W. Choi<sup>1,2,3,4</sup> | Amanda B. Zheutlin<sup>3,4</sup> | Rebecca A. Karlson<sup>3</sup> |  
Min-Jung Wang<sup>2,3</sup> | Erin C. Dunn<sup>1,3,4</sup> | Murray B. Stein<sup>5,6,7</sup> |  
Elizabeth W. Karlson<sup>8</sup> | Jordan W. Smoller<sup>1,2,3,4</sup>

<sup>1</sup>Department of Psychiatry, Massachusetts General Hospital, Boston, Massachusetts

<sup>2</sup>Department of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, Massachusetts

<sup>3</sup>Psychiatric and Neurodevelopmental Genetics Unit, Center for Genomic Medicine, Massachusetts General Hospital, Boston, Massachusetts

<sup>4</sup>Stanley Center for Psychiatric Research, Broad Institute, Boston, Massachusetts

<sup>5</sup>Department of Family Medicine and Public Health, University of California San Diego, La Jolla, California

<sup>6</sup>Department of Psychiatry, University of California San Diego, La Jolla, California

<sup>7</sup>VA San Diego Healthcare System, San Diego, California

<sup>8</sup>Department of Medicine, Division of Rheumatology, Inflammation and Immunity, Brigham and Women's Hospital, Boston, Massachusetts

## Correspondence

Karmel Choi, Ph.D., 185 Cambridge Street, Psychiatric and Neurodevelopmental Genetics Unit, Massachusetts General Hospital, Boston, MA 02114.  
Email: kwchoi@mgh.harvard.edu

## Funding information

National Human Genome Research Institute, Grant/Award Number: U01HG008685; National Institute of Mental Health, Grant/Award Number: T32MH017119

## Abstract

**Background:** Physical activity is increasingly recognized as an important modifiable factor for depression. However, the extent to which individuals with stable risk factors for depression, such as high genetic vulnerability, can benefit from the protective effects of physical activity, remains unknown. Using a longitudinal biobank cohort integrating genomic data from 7,968 individuals of European ancestry with high-dimensional electronic health records and lifestyle survey responses, we examined whether physical activity was prospectively associated with reduced risk for incident depression in the context of genetic vulnerability.

**Methods:** We identified individuals with incident episodes of depression, based on two or more diagnostic billing codes for a depressive disorder within 2 years following their lifestyle survey, and no such codes in the year prior. Polygenic risk scores were derived based on large-scale genome-wide association results for major depression. We tested main effects of physical activity and polygenic risk scores on incident depression, and effects of physical activity within stratified groups of polygenic risk.

**Results:** Polygenic risk was associated with increased odds of incident depression, and physical activity showed a protective effect of similar but opposite magnitude, even after adjusting for BMI, employment status, educational attainment, and prior depression. Higher levels of physical activity were associated with reduced odds of incident depression across all levels of genetic vulnerability, even among individuals at highest polygenic risk.

**Conclusions:** Real-world data from a large healthcare system suggest that individuals with high genetic vulnerability are more likely to avoid incident episodes of depression if they are physically active.

## KEYWORDS

depression, EHR, exercise, genomics, polygenic risk, prevention, resilience

## 1 | INTRODUCTION

Depression is a debilitating condition marked by persistent low mood, sadness, and/or loss of interest in daily activities, and represents the leading cause of disability worldwide (World Health Organization, 2018). Despite this burden, strategies to prevent depression remain limited and our understanding of robust and modifiable protective factors is incomplete. Recently, large-scale evidence (Choi et al., 2019; Schuch et al., 2018) has pointed to an important and likely causal role for physical activity—broadly defined as bodily movement resulting in energy expenditure (Caspersen, Powell, & Christenson, 1985)—in reducing the risk of depression, highlighting a promising target for prevention efforts. However, what remains unknown is the extent to which the preventive benefits of physical activity on depression may hold across different at-risk groups and even mitigate stable risk factors, such as genetic vulnerability to depression.

Depression has been observed to run within families (Lohoff, 2010) and twin studies have suggested that up to 40% of observed variation in depression can be explained by genetic differences (Sullivan, Neale, & Kendler, 2000). More recently, genome-wide association studies (GWAS) have enabled researchers to capture genetic risk for complex disorders by deriving polygenic scores that combine the effects of many variants across the genome (Wray et al., 2014). These polygenic scores have been found to explain modest though significant individual differences in outcomes such as depression (Howard et al., 2019; Wray et al., 2018). Understanding how physical activity influences depression in the context of polygenic risk for depression could allow us to better classify physical activity as a protective factor with global benefits versus one that may require consideration of individual risks, consistent with a precision psychiatry approach (Fernandes et al., 2017; Stein & Smoller, 2018). Not all interventions may be effective for all individuals, but it remains difficult to know for whom a given strategy will be efficacious. Studies suggest that people could benefit to varying degrees from preventive health measures depending on their genetic risk profiles (Klimentidis et al., 2015; Natarajan et al., 2017), but little is known regarding how physical activity could influence the risk of depression in the context of genetic factors. Such knowledge could ultimately guide clinicians and public health professionals in providing physical activity recommendations for individuals at varying levels of genetic vulnerability to depression.

Disentangling the relative influence of lifestyle and genomic factors on depression is challenging, however, to study at scale. Existing prospective studies have typically relied on actively tracked research cohorts with explicit assessments of depression (Gordon et al., 2018; Mammen & Faulkner, 2013; Schuch et al., 2018; Teychenne, Ball, & Salmon, 2008), which may not only limit sample size for comprehensive genotyping but also influence the recall and detection of depression cases. To address these issues, we capitalize on a biobank-based virtual cohort with integrated electronic health records (EHR), genomic data, and lifestyle survey responses. This genetically informed virtual cohort provides a new opportunity to

examine the prospective relationship between physical activity, genetic risk, and instances of depression occurring in naturalistic clinical data, increasing the real-world validity of this relationship. In the present study, we examined whether physical activity reduces the risk of depression in a real-world system in which diagnostic information has been documented as part of routine clinical care. Specifically, we tested the effects of physical activity on EHR-based incident depression: (a) above and beyond polygenic risk, and (b) among different groups across the spectrum of polygenic risk. We hypothesized that higher levels of physical activity would be associated with decreased odds for EHR-based incident depression even after accounting for polygenic risk, and that physical activity would demonstrate protective effects across the polygenic risk spectrum, even for individuals at high polygenic risk.

## 2 | METHODS

### 2.1 | Sample and procedures

The Partners Biobank is an ongoing virtual cohort study of patients across the Partners HealthCare hospital system (including Brigham and Women's Hospital, Massachusetts General Hospital, and other affiliated hospitals), which provides a large-scale resource of linked longitudinal EHR data, genomic data, and self-reported survey data (Karlson, Boutin, Hoffnagle, & Allen, 2016). All patients provided informed consent before enrollment, and all study procedures were approved by the Partners HealthCare Institutional Review Board.

### 2.2 | Survey measures

#### 2.2.1 | Physical activity

Participants enrolling in the Partners Biobank were invited to complete a health and lifestyle self-report survey, which assessed eight different types of recreational physical activity: walking/hiking (including commuting for work), jogging, running, biking, racquet sports, swimming, high-intensity exercise (e.g., dance and aerobics), and low-intensity exercise (e.g., yoga and stretching). Survey data were available for 11,615 Biobank participants with genomic data and physical activity responses. Participants were asked to select the average amount of time they spent per week doing each type of physical activity in the past year (see Supplementary Materials S1 for survey items and response options). To index overall levels of physical activity, average time spent per week in each activity (scaled in total hours) was summed across the eight activity types. Where a response option consisting of a range of time (e.g., 1 to < 2 hr per week) was endorsed, time spent was estimated as the midpoint of this range (i.e., 1.5 hr). Because different activity types involve varying levels of energy expenditure, we also calculated metabolic-equivalent (MET) hours of average weekly activity by taking the sum of reported time spent on each activity multiplied by their corresponding MET values (i.e., walking = 3.5, low-intensity exercise = 4.0, high-intensity exercise = 6.0, jogging = 7.0, biking = 7.0, racquet sports = 7.0, swimming = 7.0, and running = 12.0), estimated from

previous guidelines (Ainsworth et al., 1993). We observed that MET hours and total hours of physical activity were highly correlated ( $r = 0.95$ ,  $p < 2.2 \times 10^{-16}$ ). Since total hours translate more readily into actionable recommendations, we analyzed total hours as the primary analytic variable, while sensitivity analyses were performed using MET hours.

## 2.2.2 | Covariates

Demographic covariates, including participants' age and reported sex, were extracted from the lifestyle survey, as were potential confounders such as educational attainment (defined as college completion vs. not), employment status (working vs. not), and body mass index (BMI), calculated from height and weight using standard formula, that is,  $\text{kg}/\text{m}^2$ .

## 2.3 | EHR-based phenotyping

### 2.3.1 | Incident depression

By incident depression in this manuscript, we refer to evidence of incident episodes of depression (Schuch et al., 2018) rather than lifetime incident depression. We obtained comprehensive EHR data from the Research Patient Data Registry (RPDR) for Biobank participants who had lifestyle survey and genomic data ( $N = 11,615$ ), including medical encounter and diagnostic data comprising several million time-stamped billing codes from the International Classification of Diseases, 9th and 10th editions (ICD-9/10). We subset the diagnostic data to individuals who had been active in the system, defined as three or more medical encounters in their record ( $N = 11,456$ ). To establish depression incidence among individuals who were unlikely to be actively depressed at the time of their survey, we excluded those who had any depression-related billing code(s) in the year before completion of the survey (defined as any of 30 depression-related ICD-9 or ICD-10 codes corresponding to the broad Phecode 296.2 "Depression" after excluding full remission codes, as specified in the *PheWAS* R package; see Supplementary Materials S2; resulting  $N = 10,417$ ) and also retained individuals who had at least one billing code dated after completion of their lifestyle survey ( $N = 10,387$ ), indicating subsequent activity in the system.

To define case versus control status for incident depression, we identified individuals who had two or more (Wei et al., 2016) depression-related billing codes within 2 years following their lifestyle survey ( $N = 830$ ; 8%), versus those with zero codes. We selected a two-year window to capture depression episodes over a clinically relevant period of time that was relatively proximal to survey completion. To reduce the likelihood of outcome misclassification, individuals with only one code ( $N = 371$ ) were removed for analysis as in previous studies (Zheutlin et al., 2019) (resulting  $N = 10,016$ ). For sensitivity analyses, we also created a variable reflecting evidence of clinically significant depression before the study window (i.e., at any point in the system before the year before the survey).

## 2.4 | Genetic data processing

Detailed information about genotyping, quality control (QC), imputation, and population assignment procedures for the Partners Biobank is available elsewhere (Zheutlin et al., 2019). Briefly, DNA samples for each participant were genotyped on one of three Illumina arrays (MEGA, MEGA<sup>EX</sup>, and MEG BeadChip); arrays were merged and only genetic variants (i.e., single-nucleotide polymorphisms [SNPs]) on all three were retained. Variants were further excluded using filters for call rate ( $< 0.98$ ), batch effects (missingness rates differing  $> 0.01$  between any two arrays; allele frequency differences  $p < 10^{-6}$ ), and heterozygosity ( $|\text{Fhet}| > 0.2$ ). Individuals were excluded for excessive missing data ( $< 0.98$ ) or sex errors; a random individual from any pair of related individuals ( $\text{pihat} > 0.20$ ) was also excluded. We used principal components analysis to identify individuals of European ancestry based on similarity to the European population from the 1000 Genomes reference panel (The 1000 Genomes Project Consortium et al., 2015). SNPs passing this initial phase of QC were imputed and then converted to best-guess genotypes where a final round of QC was performed using genotyping probability ( $p > .80$ ), INFO score ( $> 0.90$ ), SNP missingness ( $< 0.02$ ), minor allele frequency ( $< 0.01$ ), removal of the HLA/inversion region, and violations of Hardy-Weinberg equilibrium ( $p < 10^{-6}$ ). The top 10 principal components for this sample were extracted as covariates to adjust for potential residual population stratification within the European ancestry sample.

## 2.5 | Polygenic risk scoring

We generated polygenic risk scores for depression based on publicly available summary statistics from a large GWAS meta-analysis of broad depression (discovery  $N = 500,199$  after excluding restricted 23andMe data; Howard et al., 2019). Polygenic risk scores consist of a weighted sum of each person's risk alleles at each SNP multiplied by that SNP's estimated effect size (Wray et al., 2014). Using the PRSice-2 platform (Euesden, Lewis, & O'Reilly, 2015), ambiguous SNPs were removed and SNPs were clumped using an  $r^2$  threshold of 0.20 and a 250 kb window to maximize inclusion of independent, noncorrelated SNPs. Polygenic risk scores for depression were generated based on sets of SNPs whose effects met the following  $p$ -value thresholds, in decreasing order of stringency:  $5 \times 10^{-8}$ ,  $1 \times 10^{-7}$ ,  $1 \times 10^{-6}$ ,  $1 \times 10^{-5}$ , .0001, .001, .05, .1, .5, and 1. Based on explanatory variance for the depression phenotype (Figure S3A), the  $p$ -value threshold of 1 (comprising 231,283 independent SNPs) was selected as the optimal threshold for subsequent analyses. Scores were then merged with the physical activity data, reducing the analytic sample to 7,968 individuals of European ancestry who had passed genomic QC and therefore had polygenic scores available. The top 10 principal components and array information were entered into a regression model predicting these polygenic risk scores, and residuals obtained from this model were used as stratification-adjusted polygenic risk scores in subsequent analyses.

Polygenic risk scores were approximately normally distributed across individuals (Figure S3B). As in prior studies (Choi et al., 2018; Khara et al., 2016), we divided the sample into three groups of relative polygenic risk: *low* (quintile 1;  $N = 1,594$ ), *intermediate* (quintiles 2–4;  $N = 4,781$ ), and *high* (quintile 5;  $N = 1,593$ ). As expected, these polygenic risk groups showed a dose-response relationship with incident depression (Figure S3C), with individuals at high polygenic risk showing highest odds for developing depression (OR = 1.50, 95% CI = 1.16–1.94,  $p = .002$ ), followed by those with intermediate polygenic risk (OR = 1.11, 95% CI = 0.89–1.39,  $p = .38$ ), compared with those at low polygenic risk.

## 2.6 | Statistical analyses

After standardizing the polygenic risk scores and physical activity variables (mean = 0, standard deviation (SD) = 1), we first conducted logistic regressions to examine main effects of both polygenic risk scores and physical activity levels on incident depression. Second, we tested using available data whether observed main effects were robust to potential confounding due to factors such as BMI, employment status, educational attainment, and prior depression. Third, we examined the effects of physical activity on depression within stratified groups of low, intermediate, and high polygenic risk. Finally, we explored whether main effects varied by physical activity subtype, using a Bonferroni-corrected  $p$  of .00625 that accounted for the multiple testing of eight different subtypes. All regression models were adjusted for demographic covariates and reported with profile likelihood confidence intervals.

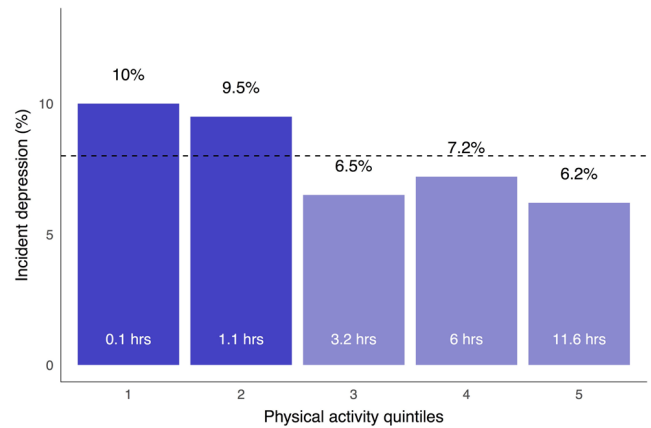
## 3 | RESULTS

### 3.1 | Sample overview and descriptive data

Participants in the analytic sample ( $N = 7,968$ ) were 57% female and aged 59.9 years on average ( $SD = 15.9$ ), with relatively high levels of educational attainment (68% with college degree or more; Supplementary Materials S4). These participants reported varying levels of overall physical activity (for distribution of scores, see Figure S4), ranging between 0 and 43 hr per week (mean = 4.4,  $SD = 4.6$ ) across all activities (mean = 21.2,  $SD = 25.6$ , for MET hours/week).

Roughly 8% ( $N = 628$ ) of the analytic sample met EHR-based criteria for incident depression within 2 years following their lifestyle survey. The two-year period prevalence of incident depression was generally higher among participants engaging in the lowest levels of physical activity (Figure 1). Descriptively, the prevalence of incident depression was above average among participants in the bottom two quintiles of physical activity, that is, those reporting about 0–1.1 hr of activity per week—and below average among participants in the upper quintiles of physical activity, that is, those reporting an average of 3.2–11.6 hr of activity per week.

As visualized in Figure 2, within each polygenic risk group, prevalence of depression was lower among individuals who engaged in higher physical activity (defined for descriptive purposes as



**FIGURE 1** Prevalence of depression per quintile of physical activity. Numbers within each bar indicate average hours of physical activity reported by participants in each quintile. Dashed line indicates the sample prevalence of incident depression (8.1%)

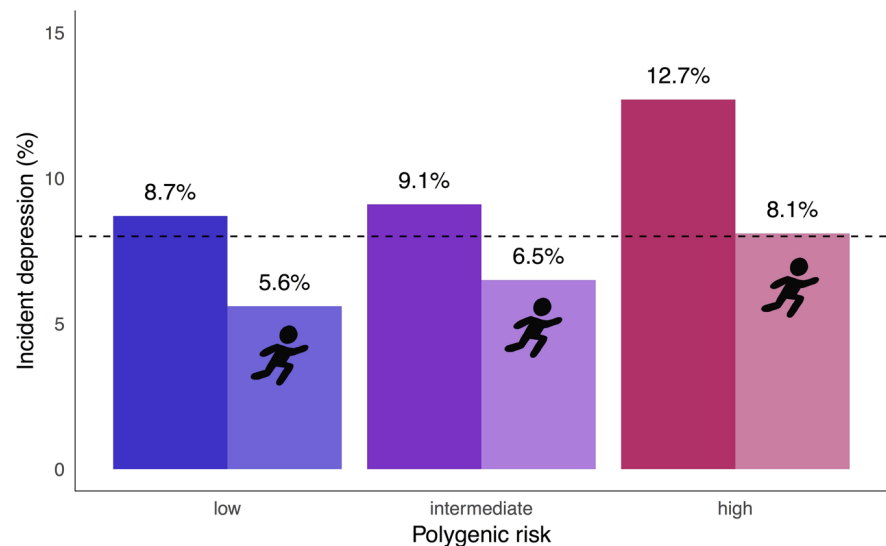
quintile 3 or higher, as per Figure 1); each of these differences was statistically significant based on tests for equality of proportions with Yates' continuity correction. For example, more active individuals with high polygenic risk showed a lower prevalence of depression (8.1%) than their inactive counterparts with similar genetic vulnerability (13%;  $p = .003$ ), though appeared relatively similar to individuals with low polygenic risk but who were less active (8.7%;  $p = .75$ ).

### 3.2 | Do polygenic risk and physical activity prospectively influence risk for incident depression?

Considered together in the same model (Supplementary Materials S6), polygenic risk scores and physical activity were each associated with incident depression. Polygenic risk scores for depression were associated with increased odds of incident depression (adjusted odds ratio (aOR) = 1.20 per 1 SD increase in polygenic risk, 95% confidence interval (CI) = 1.11–1.31,  $p = 1.04 \times 10^{-5}$ ), and higher levels of physical activity were associated with reduced odds of incident depression (aOR = 0.83 per 1 SD increase in activity, that is, roughly four extra hours per week, 95% CI = 0.75–0.90,  $p = 3.97 \times 10^{-5}$ ). The estimated effect of physical activity was virtually unchanged when considering MET hours (aOR = 0.82 per 1 SD increase in activity, 95% CI = 0.75–0.90,  $p = 5.74 \times 10^{-5}$ ). Notably, polygenic risk scores for depression were not significantly correlated with reported activity levels ( $r = 0.006$ ,  $p = .56$ ), suggesting relative specificity for incident depression.

Examining potential confounders, we found that higher BMI levels were correlated with decreased physical activity ( $r = -0.23$ ,  $p < 2.2 \times 10^{-16}$ ), increased polygenic risk for depression ( $r = 0.03$ ,  $p = .0004$ ), and incident depression ( $r = 0.04$ ,  $p = .0002$ ); conversely, higher educational attainment was correlated with increased physical activity ( $r = 0.11$ ,  $p < 2.2 \times 10^{-16}$ ), lower polygenic risk for depression ( $r = -0.03$ ,  $p = .01$ ), and reduced incident depression ( $r = -0.05$ ,  $p = 5.1 \times 10^{-6}$ ). Employment status was not correlated

**FIGURE 2** Prevalence of depression per polygenic risk group, depending on physical activity. Bar with running icon denotes group with higher physical activity (Q3–Q5; at least ~3.2 hr per week). Bar with no icon denotes group with lower physical activity (Q1–Q2; 0–1.1 hr per week). Dashed line indicates sample prevalence of incident depression (8.1%)



with incident depression ( $r = 0.01$ ,  $p = .22$ ) or polygenic risk for depression ( $r = -.005$ ,  $p = .66$ ) but was linked to increased physical activity ( $r = 0.08$ ,  $p = 9.1 \times 10^{-12}$ ). As expected, evidence of prior depression in the system was correlated with incident depression ( $r = 0.27$ ,  $p < 2.2 \times 10^{-16}$ ), polygenic risk for depression ( $r = 0.05$ ,  $p = 4.16 \times 10^{-5}$ ) and also lower activity levels at baseline ( $r = -0.04$ ,  $p = .002$ ). However, the main effects of polygenic risk and physical activity persisted even after adjusting simultaneously for BMI, educational attainment, employment status, and prior depression (aOR = 1.16 per 1SD increase in polygenic risk, 95% CI = 1.06–1.26,  $p = .0011$ ; aOR = 0.87 per 1SD increase in activity, 95% CI = 0.79–0.95,  $p = .004$ ; for full model results, see Supplementary Materials S6), suggesting that observed associations with incident depression were not primarily explained by overlaps with pre-existing BMI, employment status, educational attainment, or even prior depression.

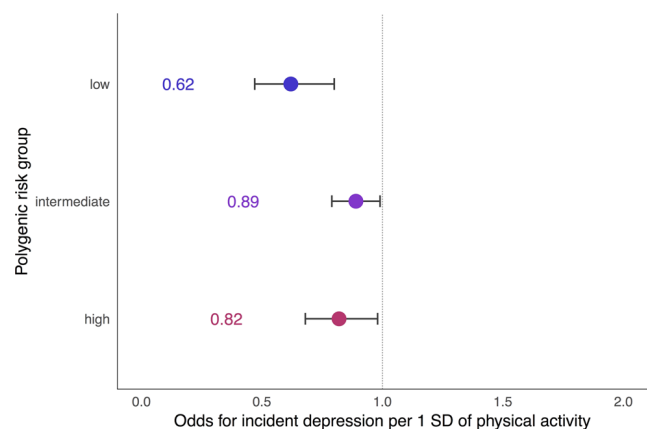
### 3.3 | What are the effects of physical activity on incident depression across the spectrum of genetic vulnerability?

Looking within groups of individuals stratified by polygenic risk (Figure 3), we observed that, compared with average levels of physical activity, engaging in higher levels of physical activity was associated with reduced odds of incident depression among individuals at low (aOR = 0.62 per 1SD increase in activity, 95% CI = 0.47–0.80,  $p = .0004$ ), intermediate (aOR = 0.89 per 1SD increase in activity, 95% CI = 0.79–0.99,  $p = .038$ ), and high polygenic risk (aOR = 0.82 per 1SD increase in activity, 95% CI = 0.68–0.98,  $p = .034$ ). This pattern of protective effects remained similar when defining polygenic risk groups based on quintiles ( $N = 2,656$  at high polygenic risk) rather than quintiles ( $N = 1,593$  at high polygenic risk), and were even more robust for the group at high polygenic risk (Supplementary Materials S5). However, overlapping confidence

intervals across all groups—and the absence of a significant multiplicative interaction effect between physical activity and polygenic risk (Supplementary Materials S6)—suggest that protective effects of physical activity are likely to hold across all levels of polygenic risk despite varying estimates within subgroups.

### 3.4 | Which physical activity subtypes showed the strongest associations with incident depression above and beyond polygenic risk?

Subtypes of physical activity were all positively correlated to a moderate extent ( $r = 0.04$ – $0.42$ ; Figure 4), with highest correlations observed between jogging and running ( $r = 0.42$ ), low-intensity exercise (e.g., yoga and stretching) and high-intensity exercise (e.g., dance, aerobic exercise, and use of exercise machine;  $r = 0.37$ ), and low-intensity exercise and walking ( $r = 0.26$ ). Examining each subtype of physical activity, we observed that running and walking each showed nominal associations with incident depression (aOR = 0.87 per 1SD increase in running, 95% CI = 0.78–0.96,  $p = .013$ ;



**FIGURE 3** Effects of physical activity on incident depression within each polygenic risk group





**FIGURE 4** Spearman correlation heatmap between reported physical activity subtypes

aOR = 0.89 per 1SD increase in walking, 95% CI = 0.80–0.97,  $p = .013$ ), even after adjusting for polygenic risk. We did not find significant associations for jogging, biking, racquet sports, or swimming. Both low-intensity exercise (e.g., yoga and stretching) and higher-intensity exercise (e.g., dance, aerobic exercise, and use of exercise machine) showed protective associations with incident depression that survived multiple testing correction (aOR = 0.86 per 1SD increase in low-intensity exercise, 95% CI = 0.77–0.95,  $p = .0062$ ; aOR = 0.84 per 1SD increase in higher-intensity exercise, 95% CI = 0.76–0.93,  $p = .0011$ ), even after adjusting for polygenic risk for depression.

## 4 | DISCUSSION

Preventing depression is a critical endeavor given its serious toll on public health. Actionable protective factors are therefore of great interest, particularly those that may offset known risk factors. Large-scale studies have supported a protective relationship between physical activity and depression (Chekroud et al., 2018; Schuch et al., 2018) and recent evidence suggests this relationship is broadly causal (Choi et al., 2019), however, it has remained unclear whether the beneficial effects of physical activity apply in the presence of genetic vulnerability to depression, or whether genetic factors attenuate the protective influences of physical activity. Using a longitudinal biobank-based virtual cohort integrating genomic data with EHR and lifestyle survey responses ( $N = 7,968$ ), we investigated whether physical activity is prospectively linked to reduced risk for depression even in the context of genetic vulnerability. The most striking finding from this study was that, when considered along with polygenic risk (a putatively fixed biological factor), physical activity (a modifiable behavioral factor) showed comparable effects on incident depression but in an opposite, protective direction. This suggests that while stable genetic factors shape risk for depression, potent modifiable avenues for prevention also exist.

Using clinical EHR data, we observed that higher levels of physical activity at baseline were linked to significantly reduced chances of appearing in the healthcare system with a depressive disorder over the next 2 years. Descriptive findings indicated that individuals engaging in three or more hours of activity per week showed reduced prevalence of incident depression. Regression models indicated that individuals experienced a 17% decrease in the odds of incident depression for each 1SD increase in reported activity, roughly equivalent to four extra hours of activity per week. Together, this suggests that approximately 45 additional minutes of physical activity each day could translate to meaningful reductions in a person's risk for depression. This estimate is broadly consistent with previous work (Choi et al., 2019; Schuch et al., 2018) but the real-world context of health records represents an important step in translating known associations between physical activity and depression into the clinic.

Adding to emerging psychiatric research integrating genetic and EHR data (Musliner et al., 2019), we found that polygenic risk for depression was prospectively associated with depression outcomes in a healthcare system, such that individuals with the highest polygenic loading for depression showed over 50% higher odds of incident depression in a two-year window than those at low polygenic risk. This association persisted even when accounting for EHR evidence of prior depression, suggesting that polygenic risk scores have potential not only to explain lifetime depression but also the likelihood of incident episodes—though remains to be further tested.

Looking within polygenic risk groups, physical activity showed protective effects in all groups, with the largest effect observed among individuals with lowest polygenic risk for depression. Qualitative differences should be interpreted with caution, however, as the overall pattern of results (e.g., overlapping confidence intervals and absence of a significant modifying effect) indicates that depression-related benefits of physical activity were not significantly different across the spectrum of polygenic risk. Our findings point to

two complementary insights. On one hand, individuals at high genetic risk for depression may indeed benefit to a similar degree from physical activity. On the other hand, their baseline (genetic) vulnerability to depression is nonetheless elevated; being more active could help them resemble individuals at lower genetic risk who are inactive, but not quite as much as those at lower genetic risk who were also active (Figure 2). As such, additional exercise and/or other protective factors may still be needed for individuals at high genetic risk to reduce overall depression risk to the same level we would expect to see among those without such pre-existing vulnerabilities.

Interestingly, we observed that overall time spent across any activity was just as good if not better than time adjusted for the intensity of activity (i.e., MET hours) in explaining risk for depression, and these metrics were closely correlated. In exploring subtypes of physical activity, we saw that both higher-intensity forms of activity such as aerobic exercise, dance, and/or use of an exercise machine, and lower-intensity exercise such as yoga and stretching, were strongly inversely related to incident depression. Running and walking also showed protective associations with depression but were less robust on their own. These findings align with previous research suggesting a mood-enhancing role of aerobic exercise (Chekroud et al., 2018; Schuch et al., 2016) as well as benefits of lighter activities such as walking (Mammen & Faulkner, 2013), though our findings support the broader conclusion that cumulative physical activity of any kind has a positive impact on depression risk even in the presence of genetic factors.

#### 4.1 | Limitations

This study had a number of methodological limitations. First, this prospective work was observational and noninterventive, and therefore does not reveal causal relationships between physical activity and risk of depression, which have been tested using other methodological approaches (Choi et al., 2019). Second, despite their advantages, EHR-based diagnostic codes are provider-based and may not capture nuanced aspects of a depressive episode such as severity or course. In addition, there may be instances of clinical care occurring outside of the hospital that are not captured by the billing codes within our system, though we sought to only include individuals who remained active in the system following their lifestyle survey. While our depression phenotype may be relatively noisy, it represents a real-world index of clinically significant depression and it is notable that we nonetheless detected genetic signals for this EHR-based depression phenotype. Third, we studied incident depression episodes (i.e., transitioning to active depression) rather than new lifetime onset depression, given the episodic nature of depressive disorders and because lifetime history may not be captured in the span of available EHR data. It is possible that some genetic influences on pre-existing depression may partially explain the observed prospective relationships between physical activity and incident depression. While we ruled out any obvious correlation between polygenic risk for depression and reported physical activity levels, there could be subthreshold genetic factors common to

depression and physical activity that influenced these relationships. However, it is notable that our findings persisted even after adjusting for evidence of prior depression. Fourth, caution should be taken with generalizability as this was a highly educated sample that consented to EHR-based research and only included individuals of European ancestry since the effect sizes from the discovery GWAS for depression were also generated in a sample of European ancestry. Lastly, we used a measure of self-reported activity reflecting average levels in the past year, while true levels of physical activity may fluctuate over the course of a year. This measure did not capture all forms of physical movement (e.g., incidental activity such as using stairs or doing chores, or work-related exertion such as standing, walking, bending, or lifting) that could yield preventive effects on depression risk (Stamatakis et al., 2019); the observed effects of physical activity may have been even stronger if these forms were included, and future studies leveraging objective modes of activity tracking can help obviate these challenges. In our recent Mendelian randomization study (Choi et al., 2019), only objectively measured physical activity showed an association with depression risk. While this indicates that effects of self-reported activity on depression cannot be verified using current genetic instruments in a Mendelian randomization framework, it does not rule out self-reported activity as a useful proxy of physical activity relevant for depression. Self-reported activity may also be subject to mood-related biases (e.g., negative memory and self-evaluation) that underlie depression risk (Stamatakis et al., 2019), although adjusting for genetic influences on depression may indirectly address this issue and represents a novel strategy for controlling for these influences, to some extent.

#### 5 | CONCLUSION

Although depression is the leading cause of disability worldwide, there are few established protective factors for depression and limited knowledge of how broadly applicable they might be. We provide real-world evidence of the prospective impact of physical activity in mitigating risk of depression even in the context of genetic vulnerability. In terms of clinical implications, individuals may believe that static risk factors (e.g., genetic risk and family history) are deterministic and may be unsure whether engaging in preventive behaviors such as physical activity will make a difference. The results of this study have potential to aid primary care and mental health providers in combating these perceptions by providing empirical evidence that physical activity can alter real-world depression risk even among individuals at high genetic vulnerability, and can therefore guide counseling and first-line recommendations for individuals who have a substantial family history of depression or who may, in the future, receive more detailed information about their genomic risk. As the field increasingly quantifies the role of genome-wide factors in complex conditions such as depression (Howard et al., 2019; Wray et al., 2018), providing people with specific knowledge that a modifiable factor such as physical activity can exert protective

effects above and beyond the influence of one's genetic makeup can enrich counseling efforts and potentially motivate shifts in lifestyle behavior to improve mental health.

## ACKNOWLEDGMENTS

We would like to thank the Partners Biobank and its participants for providing samples, genomic data, and health information data. K. W. C. was supported in part by a NIMH T32 Training Fellowship (T32MH017119). J. W. S is a Tepper Family MGH Research Scholar and supported in part by the Demarest Lloyd, Jr, Foundation and an NHGRI grant supporting the eMERGE Network (U01HG008685).

## CONFLICT OF INTERESTS

Dr. Stein has in the past 3 years been a consultant for Actelion, Aptinyx, Janssen, Jazz Pharmaceuticals, Neurocrine Biosciences, Oxeia Biopharmaceuticals, and Pfizer. Dr. Stein has stock options in Oxeia Biopharmaceuticals. Dr. Smoller is an unpaid member of the Scientific Advisory Board of Psy Therapeutics, Inc and of the Bipolar/Depression Research Community Advisory Panel of 23andMe. Dr. Zheutlin has received salary for work at Sema4 and Dr. Wang for work at Analysis Group. The other authors do not have potential interests to declare.

## DATA AVAILABILITY STATEMENT

Data that support the findings of this study are available from the Partners Biobank (for more information, see: <https://biobank.partners.org/for-researchers>). Restrictions apply to the availability of these data, which are available to Partners-affiliated researchers via formal application to the Partners Biobank.

## ORCID

Karmel W. Choi  <http://orcid.org/0000-0002-3914-2431>

Min-Jung Wang  <http://orcid.org/0000-0003-4432-3330>

Erin C. Dunn  <http://orcid.org/0000-0003-1413-3229>

Murray B. Stein  <http://orcid.org/0000-0001-9564-2871>

Jordan W. Smoller  <http://orcid.org/0000-0002-0381-6334>

## REFERENCES

- Ainsworth, B. E., Haskell, W. L., Leon, A. S., Jacobs, D. R., Montoye, H. J., Sallis, J. F., & Paffenbarger, R. S. (1993). Compendium of physical activities: Classification of energy costs of human physical activities. *Medicine & Science in Sports & Exercise*, 25(1), 71–80.
- Caspersen, C. J., Powell, K. E., & Christenson, G. M. (1985). Physical activity, exercise, and physical fitness: Definitions and distinctions for health-related research. *Public Health Reports (Washington, D.C.: 1974)*, 100(2), 126–131.
- Chekroud, S. R., Gueorguieva, R., Zheutlin, A. B., Paulus, M., Krumholz, H. M., Krystal, J. H., & Chekroud, A. M. (2018). Association between physical exercise and mental health in 1.2 million individuals in the USA between 2011 and 2015: A cross-sectional study. *The Lancet Psychiatry*, 5(9), 739–746. [https://doi.org/10.1016/S2215-0366\(18\)30227-X](https://doi.org/10.1016/S2215-0366(18)30227-X)
- Choi, K. W., Chen, C.-Y., Stein, M. B., Klimentidis, Y. C., Wang, M. -J., Koenen, K. C., ... for the Major Depressive Disorder Working Group of the Psychiatric Genomics Consortium (2019). Assessment of bidirectional relationships between physical activity and depression among adults: A 2-sample mendelian randomization study. *JAMA Psychiatry*, <https://doi.org/10.1001/jamapsychiatry.2018.4175>
- Choi, K. W., Chen, C.-Y., Sun, X., Jain, S., Kessler, R., Koenen, K., ... Smoller, J. (2018). Prospective study of polygenic risk, protective factors, and incident depression following combat deployment in US army soldiers. *Psychological Medicine*, 15(4), 1–9.
- Euesden, J., Lewis, C. M., & O'Reilly, P. F. (2015). PRSice: Polygenic risk score software. *Bioinformatics (Oxford, England)*, 31(9), 1466–1468. <https://doi.org/10.1093/bioinformatics/btu848>
- Fernandes, B. S., Williams, L. M., Steiner, J., Leboyer, M., Carvalho, A. F., & Berk, M. (2017). The new field of 'precision psychiatry'. *BMC Medicine*, 15(1), 80. <https://doi.org/10.1186/s12916-017-0849-x>
- Gordon, B. R., McDowell, C. P., Hallgren, M., Meyer, J. D., Lyons, M., & Herring, M. P. (2018). Association of efficacy of resistance exercise training with depressive symptoms: Meta-analysis and meta-regression analysis of randomized clinical trials. *JAMA Psychiatry*, 75, 566–576. <https://doi.org/10.1001/jamapsychiatry.2018.0572>
- Howard, D. M., Adams, M. J., Clarke, T. K., Hafferty, J. D., Gibson, J., Shirali, M., ... Coleman, J. R. I. (2019). Genome-wide meta-analysis of depression identifies 102 independent variants and highlights the importance of the prefrontal brain regions. *Nature Neuroscience*, 22(3), 343–352. <https://doi.org/10.1038/s41593-018-0326-7>
- Karlson, E., Boutin, N., Hoffnagle, A., & Allen, N. (2016). Building the partners healthcare biobank at partners personalized medicine: Informed consent, return of research results, recruitment lessons and operational considerations. *Journal of Personalized Medicine*, 6(1), 2. <https://doi.org/10.3390/jpm6010002>
- Khera, A. V., Emdin, C. A., Drake, I., Natarajan, P., Bick, A. G., Cook, N. R., ... Chasman, D. I. (2016). Genetic risk, adherence to a healthy lifestyle, and coronary disease. *New England Journal of Medicine*, 375(24), 2349–2358. <https://doi.org/10.1056/NEJMoa1605086>
- Klimentidis, Y. C., Bea, J. W., Lohman, T., Hsieh, P. -S., Goings, S., & Chen, Z. (2015). High genetic risk individuals benefit less from resistance exercise intervention. *International Journal of Obesity*, 39(9), 1371–1375. <https://doi.org/10.1038/ijo.2015.78>
- Lohoff, F. W. (2010). Overview of the genetics of major depressive disorder. *Current Psychiatry Reports*, 12(6), 539–546. <https://doi.org/10.1007/s11920-010-0150-6>
- Mammen, G., & Faulkner, G. (2013). Physical activity and the prevention of depression. *American Journal of Preventive Medicine*, 45(5), 649–657. <https://doi.org/10.1016/j.amepre.2013.08.001>
- Musliner, K. L., Mortensen, P. B., McGrath, J. J., Suppli, N. P., Hougaard, D. M., Bybjerg-Grauholm, J., ... Bækvad-Hansen, M. (2019). Association of polygenic liabilities for major depression, bipolar disorder, and schizophrenia with risk for depression in the Danish population. *JAMA Psychiatry*, 76, 516. <https://doi.org/10.1001/jamapsychiatry.2018.4166>
- Natarajan, P., Young, R., Stitzel, N. O., Padmanabhan, S., Baber, U., Mehran, R., ... Sartori, S. (2017). Polygenic risk score identifies subgroup with higher burden of atherosclerosis and greater relative benefit from statin therapy in the primary prevention setting. *Circulation*, 135(22), 2091–2101. <https://doi.org/10.1161/CIRCULATIONAHA.116.024436>
- Schuch, F. B., Vancampfort, D., Firth, J., Rosenbaum, S., Ward, P. B., Silva, E. S., ... Stubbs, B. (2018). Physical activity and incident depression: A meta-analysis of prospective cohort studies. *American Journal of Psychiatry*, 175(7), 631–648. <https://doi.org/10.1176/appi.ajp.2018.17111194>
- Schuch, F. B., Vancampfort, D., Richards, J., Rosenbaum, S., Ward, P. B., & Stubbs, B. (2016). Exercise as a treatment for depression: A



- meta-analysis adjusting for publication bias. *Journal of Psychiatric Research*, 77, 42–51. <https://doi.org/10.1016/j.jpsychires.2016.02.023>
- Stamatakis, E., Johnson, N. A., Powell, L., Hamer, M., Rangul, V., & Holtermann, A. (2019). Short and sporadic bouts in the 2018 US physical activity guidelines: Is high-intensity incidental physical activity the new HIIT? *British Journal of Sports Medicine*, 53, 1137–1139. <https://doi.org/10.1136/bjsports-2018-100397>
- Stein, M. B., & Smoller, J. W. (2018). Precision psychiatry—will genomic medicine lead the way? *JAMA Psychiatry*, 75(7), 663. <https://doi.org/10.1001/jamapsychiatry.2018.0375>
- Sullivan, P. F., Neale, M. C., & Kendler, K. S. (2000). Genetic epidemiology of major depression: Review and meta-analysis. *American Journal of Psychiatry*, 157(10), 1552–1562. <https://doi.org/10.1176/appi.ajp.157.10.1552>
- Teychenne, M., Ball, K., & Salmon, J. (2008). Physical activity and likelihood of depression in adults: A review. *Preventive Medicine*, 46(5), 397–411. <https://doi.org/10.1016/j.ypmed.2008.01.009>
- The 1000 Genomes Project Consortium, Gibbs, R. A., Boerwinkle, E., Doddapaneni, H., Han, Y., Korchina, V., ... Rasheed, A. (2015). A global reference for human genetic variation. *Nature*, 526(7571), 68–74. <https://doi.org/10.1038/nature15393>
- Wei, W. -Q., Teixeira, P. L., Mo, H., Cronin, R. M., Warner, J. L., & Denny, J. C. (2016). Combining billing codes, clinical notes, and medications from electronic health records provides superior phenotyping performance. *Journal of the American Medical Informatics Association*, 23(e1), e20–e27. <https://doi.org/10.1093/jamia/ocv130>
- World Health Organization. (2018, March 22). Depression. Retrieved from Fact sheets website: <https://www.who.int/news-room/fact-sheets/detail/depression>
- Wray, N. R., Lee, S. H., Mehta, D., Vinkhuyzen, A. A. E., Dudbridge, F., & Middeldorp, C. M. (2014). Research review: Polygenic methods and their application to psychiatric traits. *Journal of Child Psychology and Psychiatry*, 55(10), 1068–1087. <https://doi.org/10.1111/jcpp.12295>
- Wray, N. R., Ripke, S., Mattheisen, M., Trzaskowski, M., Byrne, E. M., Abdellaoui, A., ... Adams, M. J. (2018). Genome-wide association analyses identify 44 risk variants and refine the genetic architecture of major depression. *Nature Genetics*, 50(5), 668–681. <https://doi.org/10.1038/s41588-018-0090-3>
- Zheutlin, A. B., Dennis, J., Linnér, R. K., Moscati, A., Restrepo, N., Straub, P., ... Smoller, J. W. (2019). Penetrance and pleiotropy of polygenic risk scores for schizophrenia in 106,160 patients across four healthcare systems. *BioRxiv*, <https://doi.org/10.1101/421164>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section.

**How to cite this article:** Choi KW, Zheutlin AB, Karlson RA, et al. Physical activity offsets genetic risk for incident depression assessed via electronic health records in a biobank cohort study. *Depress Anxiety*. 2019;1–9. <https://doi.org/10.1002/da.22967>

## Supplementary Materials

### S1. Physical Activity Survey Items

*During the past year, what was your average time per week spent at each of the following recreational activities?*

1. Walking or hiking (including to/from/for work)
2. Jogging (slower than 10 mins/mile)
3. Running (10 mins/mile or faster)
4. Bicycling (including stationary machine)
5. Tennis, squash or racquetball
6. Lap swimming
7. Higher intensity exercise (aerobic exercise, aerobic dance, exercise machine)
8. Lower intensity exercise (yoga, stretching, toning)

Response options: *None (0 mins per week); 1-19 mins per week; 20- 59 mins per week; 1 hour per week; 1-<2 hours per week; 2-<3 hours per week; 3-<6 hours per week; 6+ hours per week*

Source: Partners Biobank Health Behaviors & Family History Survey

## S2. EHR-based Phenotyping

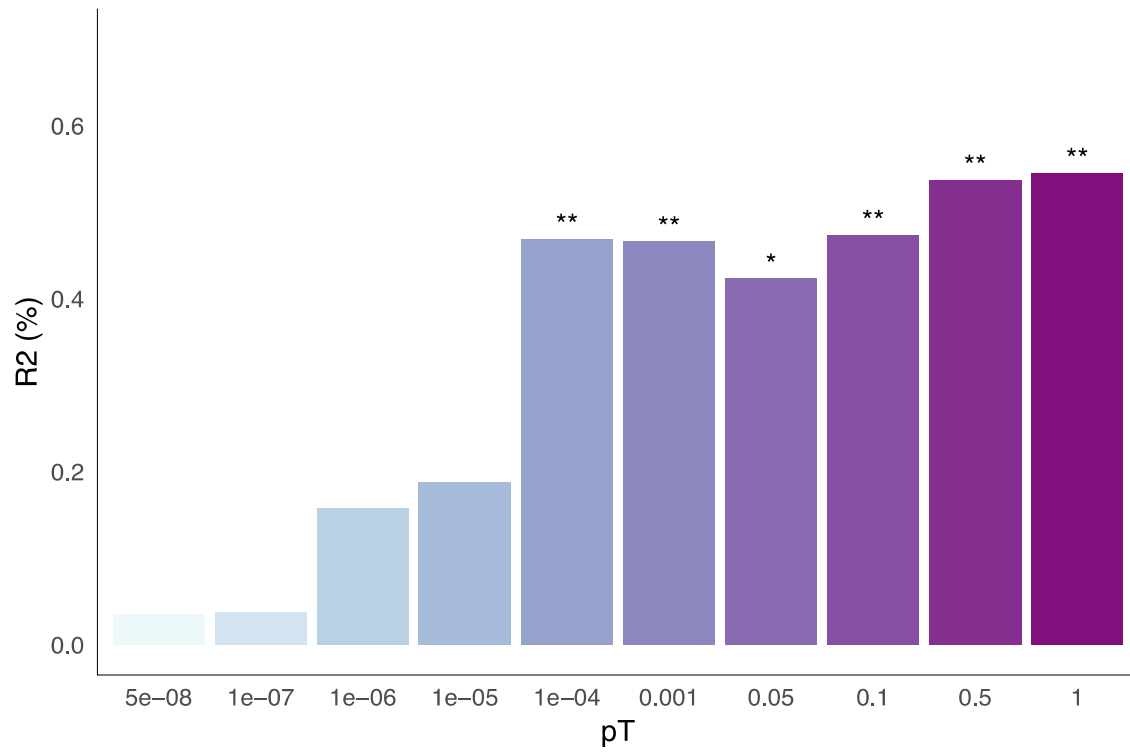
**Table S2. Depression-related ICD-9 and 10 codes mapped to Phecode 296.2 (“Depression”)**

ICD	Code	
9	296.20	Major depressive affective disorder, single episode, unspecified
9	296.25	Major depressive affective disorder, single episode, in partial or unspecified remission
9	296.35	Major depressive affective disorder, recurrent episode, in partial or unspecified remission
9	296.2	Major depressive disorder single episode
9	296.22	Major depressive affective disorder, single episode, moderate
9	296.3	Major depressive disorder recurrent episode
9	296.24	Major depressive affective disorder, single episode, severe, specified as with psychotic behavior
9	311	Depressive disorder, not elsewhere classified
9	296.31	Major depressive affective disorder, recurrent episode, mild
9	296.34	Major depressive affective disorder, recurrent episode, severe, specified as with psychotic behavior
9	296.23	Major depressive affective disorder, single episode, severe, without mention of psychotic behavior
9	296.33	Major depressive affective disorder, recurrent episode, severe, without mention of psychotic behavior
9	296.30	Major depressive affective disorder, recurrent episode, unspecified
9	296.21	Major depressive affective disorder, single episode, mild
9	296.32	Major depressive affective disorder, recurrent episode, moderate
10	F32	Major depressive disorder, single episode
10	F32.0	Major depressive disorder, single episode, mild
10	F32.1	Major depressive disorder, single episode, moderate
10	F32.2	Major depressive disorder, single episode, severe without psychotic features
10	F32.3	Major depressive disorder, single episode, severe with psychotic features
10	F32.4	Major depressive disorder, single episode, in partial remission
10	F32.9	Major depressive disorder, single episode, unspecified
10	F33	Major depressive disorder, recurrent
10	F33.0	Major depressive disorder, recurrent, mild
10	F33.1	Major depressive disorder, recurrent, moderate
10	F33.2	Major depressive disorder, recurrent severe without psychotic features
10	F33.3	Major depressive disorder, recurrent, severe with psychotic symptoms
10	F33.40	Major depressive disorder, recurrent, in remission, unspecified
10	F33.41	Major depressive disorder, recurrent, in partial remission
10	F33.9	Major depressive disorder, recurrent, unspecified

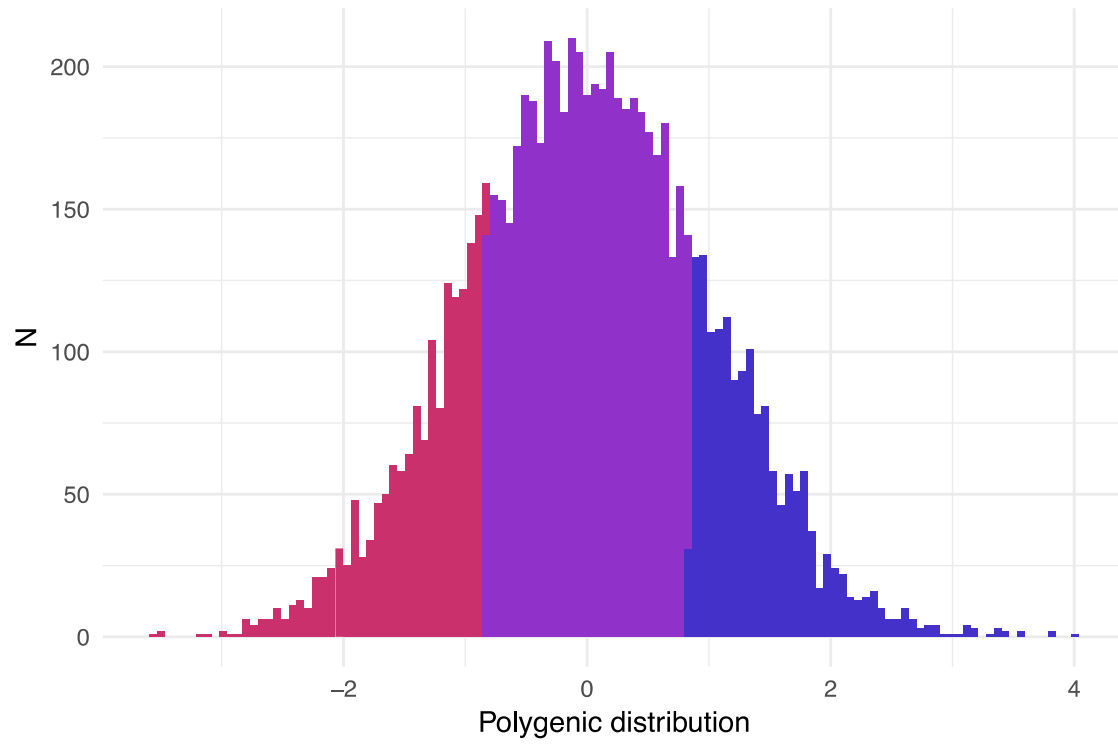
Excluding 296.26, 296.36, F32.5, F33.42 codes (“in full remission”)

### S3. Polygenic Risk Scoring

**Figure S3A. Assessing the explanatory power of polygenic risk scores based on different p-value thresholds (pT), adjusting for age, sex, top 10 PCs, and batch effects**

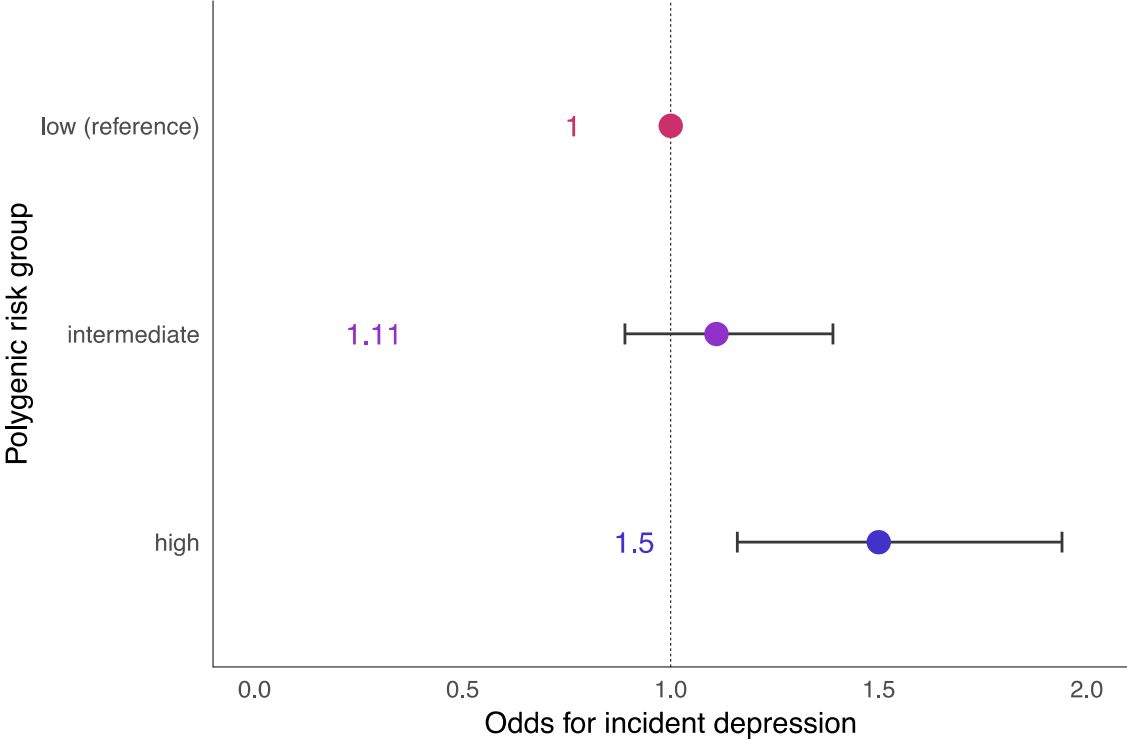


**Figure S3B. Polygenic risk distribution, divided into low (quintile 1), intermediate (quintile 2-4), and high risk (quintile 5) groups**





**Figure S3C. Dose-response effect of polygenic risk groups on incident depression**



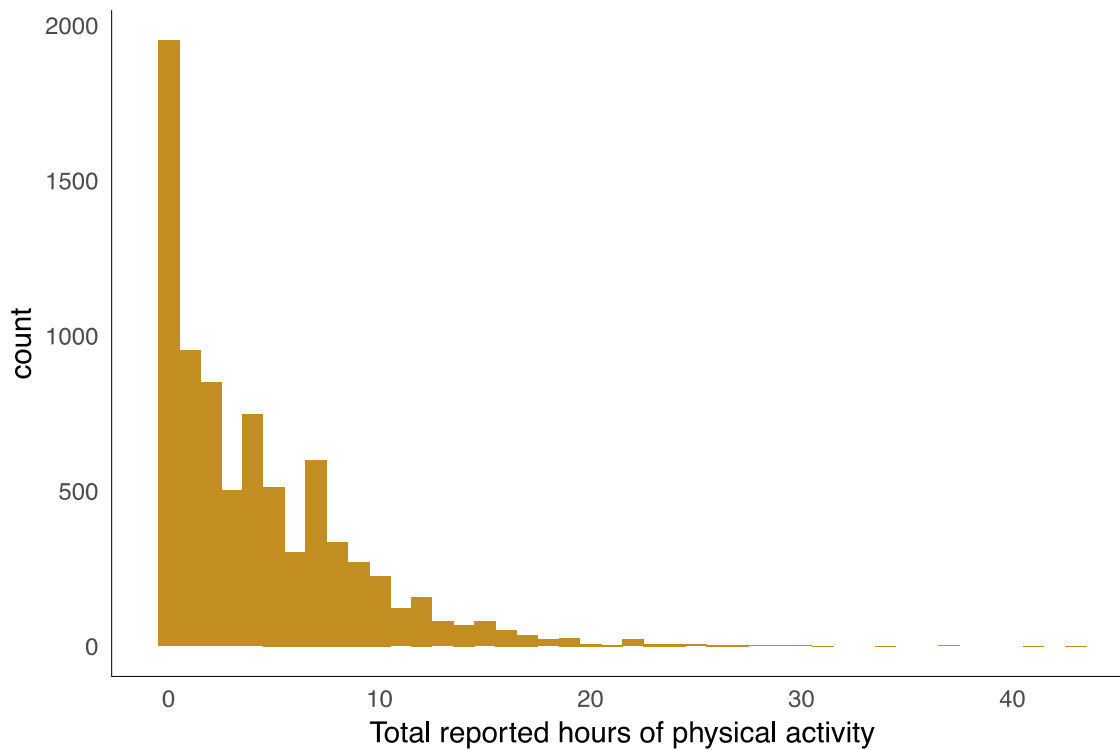
#### S4. Descriptive data

**Table S4. Descriptive characteristics summary table**

	N (%)	M (SD)
Met criteria for incident depression	628 (8%)	
Physical activity (hours/week)		4.4 (4.6)
Physical activity (MET hours/week)		22.2 (25.6)
Female <sup>a</sup>	4567 (57%)	
Age		59.9 (15.9)
BMI <sup>b</sup>		27.5 (6.0)
Completed higher education <sup>c</sup>	5266 (68%)	
Employed <sup>d</sup>	5592 (73%)	
Met criteria for any prior depression	619 (8%)	

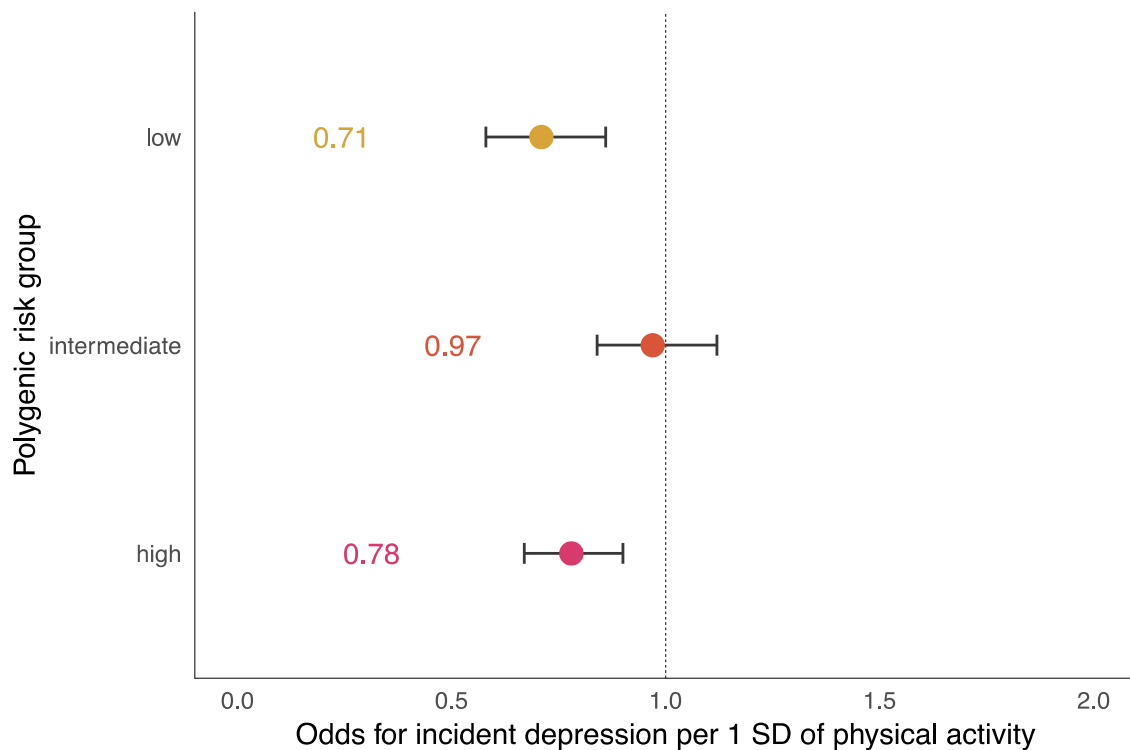
<sup>a</sup> n=7961; <sup>b</sup> n=7698; <sup>c</sup> n=7695; <sup>d</sup> n=7719 with available data

**Figure S4. Distribution of physical activity levels in the sample**



## S5. Sensitivity analyses: Defining polygenic risk groups based on tertiles

To evaluate the impact of sample size for detecting effects at the ends of the polygenic risk spectrum, we divided our sample into equal tertiles of polygenic risk ( $N = 2,656$  in each tertile) instead of using the top and bottom quintiles to define high and low polygenic risk ( $N=1,593$  and  $1,594$ , respectively). Interestingly, similar to quintile-based estimates, we found that individuals with high polygenic risk based on tertiles showed comparably elevated odds for incident depression (aOR = 1.50, 95% CI = 1.23-1.84,  $p = 0.000078$ ), followed by those with intermediate polygenic risk (aOR = 1.16, 95% CI = 0.94-1.43,  $p = 0.18$ )—suggesting that polygenic influences on depression in our sample may not be concentrated among those with very high genetic risk but rather exist more broadly in the upper range of the distribution. Within polygenic risk tertiles, we observed the effect of physical activity remained strongest for those at low polygenic risk (aOR = 0.71, 95% CI = 0.58-0.86,  $p=0.00058$ ) but remained significant those at high polygenic risk (aOR = 0.78, 95% CI = 0.67-0.90,  $p = 0.00096$ ), while being the weakest for those at intermediate polygenic risk (aOR = 0.97, 95% CI = 0.84-1.12,  $p = 0.70$ ) (**Figure S5**). However, overlapping confidence intervals suggested broadly consistent effects of physical activity across all groups of polygenic risk.



**Figure S5. Effect of physical activity on incident depression within each tertile-based polygenic risk group**

## S6. Logistic regression models with formal tests of multiplicative interaction and covariates

**Table S6A. Polygenic risk, physical activity, and incident MDD**

Variable	Model 1		Model 2 (interaction)	
	aOR (95%CI)	p-value	aOR (95%CI)	p-value
Polygenic risk	1.20 (1.11-1.31)	1.04 x 10 <sup>-5</sup>	1.21 (1.11-1.31)	6.70 x 10 <sup>-6</sup>
Physical activity	0.83 (0.75-0.90)	3.97 x 10 <sup>-5</sup>	0.82 (0.74-0.90)	2.42 x 10 <sup>-5</sup>
Age	0.98 (0.97-0.98)	1.26 x 10 <sup>-16</sup>	0.98 (0.97-0.98)	1.22 x 10 <sup>-16</sup>
Sex (male)	0.74 (0.62-0.88)	5.88 x 10 <sup>-4</sup>	0.73 (0.62-0.87)	5.62 x 10 <sup>-4</sup>
Polygenic risk*activity	-	-	1.05 (0.96-1.15)	0.27

aOR=adjusted odds ratio. 95% CI=95% confidence interval. N=7,968

**Table S6B. Adjusted models of polygenic risk, physical activity, and incident MDD**

Variable	Model 1		Model 2 (with prior depression)	
	aOR (95%CI)	p-value	aOR (95%CI)	p-value
Polygenic risk	1.20 (1.10-1.30)	2.38 x 10 <sup>-5</sup>	1.16 (1.06-1.26)	1.09 x 10 <sup>-3</sup>
Physical activity	0.86 (0.78-0.94)	1.32 x 10 <sup>-3</sup>	0.87 (0.79-0.95)	3.55 x 10 <sup>-3</sup>
Age	0.98 (0.97-0.98)	4.65 x 10 <sup>-17</sup>	0.97 (0.97-0.98)	8.20 x 10 <sup>-19</sup>
Sex (male)	0.74 (0.61-0.88)	7.84 x 10 <sup>-4</sup>	0.86 (0.71-1.04)	0.11
BMI	1.02 (1.01-1.03)	5.64 x 10 <sup>-3</sup>	1.01 (0.998-1.03)	0.10
Higher education	0.70 (0.58-0.83)	5.95 x 10 <sup>-5</sup>	0.68 (0.57-0.82)	4.64 x 10 <sup>-5</sup>
Employed	0.89 (0.73-1.10)	0.27	0.96 (0.77-1.19)	0.67
Any prior depression	-	-	8.06 (6.56-9.88)	7.96 x 10 <sup>-89</sup>

aOR=adjusted odds ratio. 95% CI=95% confidence interval. N=7,675 with available data