

UCLA

UCLA Previously Published Works

Title

Exploring the multiverse: the impact of researchers analytic decisions on relationships between depression and inflammatory markers.

Permalink

<https://escholarship.org/uc/item/2596z95b>

Journal

Neuropsychopharmacology, 48(10)

Authors

Rengasamy, Manivel

Moriarity, Daniel

Kraynak, Thomas

et al.

Publication Date

2023-09-01

DOI

10.1038/s41386-023-01621-4

Peer reviewed

ARTICLE



Exploring the multiverse: the impact of researchers' analytic decisions on relationships between depression and inflammatory markers

Manivel Rengasamy¹✉, Daniel Moriarity^{2,3}, Thomas Kraynak^{4,5}, Brenden Tervo-Clemmens⁶ and Rebecca Price^{1,4}

© The Author(s), under exclusive licence to American College of Neuropsychopharmacology 2023

In recent years, a replication crisis in psychiatry has led to a growing focus on the impact of researchers' analytic decisions on the results from studies. Multiverse analyses involve examining results across a wide array of possible analytic decisions (e.g., log-transforming variables, number of covariates, or treatment of outliers) and identifying if study results are robust to researchers' analytic decisions. Studies have begun to use multiverse analysis for well-studied relationships that have some heterogeneity in results/conclusions across studies. We examine the well-studied relationship between peripheral inflammatory markers (PIMs; e.g., white blood cell count (WBC) and C-reactive protein (CRP)) and depression severity in the large NHANES dataset ($n = 25,962$). Specification curve analyses tested the impact of 9 common analytic decisions (comprising of 58,000+ possible combinations) on the association of PIMs and depression severity. Relationships of PIMs and total depression severity are robust to analytic decisions (based on tests of inference jointly examining effect sizes and p -values). However, moderate/large differences are noted in effect sizes based on analytic decisions and the majority of analyses do not result in significant findings, with the percentage of analyses with statistically significant results being 46.1% for WBC and 43.8% for CRP. For associations of PIMs with specific symptoms of depression, some associations (e.g., sleep, appetite) in males (but not females) were robust to analytic decisions. We discuss how multiverse analyses can be used to guide research and also the need for authors, reviewers, and editors to incorporate multiverse analyses to enhance replicability of research findings.

Neuropsychopharmacology (2023) 48:1465–1474; <https://doi.org/10.1038/s41386-023-01621-4>

INTRODUCTION

Over the past decade, researchers have become increasingly oriented towards the reproducibility of results, given that a wide range of research findings have failed to replicate [1]. The reproducibility of results is critical because it forms the backbone of the scientific method in generating hypotheses and ultimately informs clinical care. Though most researchers are aware of broad challenges to reproducibility, identifying the degree to which any specific result may be reproducible has been difficult. Given these concerns, some scientific fields have begun to investigate one critical factor that influences reproducibility – the effects of variability in *researchers' analytic decisions* on the results of analyses (related to “researcher's degrees of freedom” and the “garden of forking paths”) [2, 3]. One rigorous technique, called “multiverse analyses”, has been developed to assess effects of analytic decisions [4]. Multiverse analyses have found that researchers' analytic decisions, such as variable transformation (e.g., log transformation) or exclusion of certain participants, often lead to discordant support for a single hypothesis even when the exact same dataset is utilized [5]. Though strategies to improve reproducibility (e.g., pre-registry, meta-analyses, and data transparency) exist, such strategies are still vulnerable to the choice of

analytic decisions made by researchers. These researcher-level analytic choices may lead to fragility of results – namely, failure to correctly reject the null hypothesis, incorrectly rejecting the null hypothesis, or incorrect effect size estimation.

At its core, multiverse analyses involve testing numerous combinations of common and appropriate analytic decisions. Given that a wide range of analytic pipelines can be justified and defensible by knowledgeable researchers, multiverse analyses provide a methodology to quantify uncertainty and account for the “researcher degrees of freedom” that can better determine the probability that a researcher using any of a variety of methods would obtain a similar result from the exact same dataset. Multiverse analyses can also be used to identify *which* analytic techniques (e.g., excluding individuals with comorbid illnesses) might result in most (or least) replicable associations. This information can inform ideal methodology for use in future studies. Furthermore, recognition of the *type* of analytic decisions that cause variance in reproducibility may spark further studies. For instance, if differences in reproducibility exist in participants based on inclusion of outlier values of a variable, then further research could explore the reasons why outliers of a given variable affect results.

¹Department of Psychiatry, University of Pittsburgh, Pittsburgh, PA, USA. ²Semel Institute for Neuroscience and Human Behavior, University of California Los Angeles, Los Angeles, CA, USA. ³Department of Genetics, Stanford University, Stanford, CA, USA. ⁴Department of Psychology, University of Pittsburgh, Pittsburgh, PA, USA. ⁵Department of Epidemiology, University of Pittsburgh, Pittsburgh, PA, USA. ⁶Center for Addiction Medicine, Harvard Medical School, Boston, MA, USA. ✉email: rengasamym@upmc.edu

Received: 12 January 2023 Revised: 28 April 2023 Accepted: 23 May 2023
Published online: 19 June 2023

One area of particular interest in the field of behavioral medicine with regards to reproducibility is the relationship between peripheral inflammatory markers (PIMs; e.g., C-reactive protein (CRP) and white blood count (WBC)) and depression (defined by both total depressive symptom severity and at the level of each depressive symptom), given direct clinical implications for treatment of depression (or risk factors for depression) and indirect impact of such knowledge on broader health outcomes linked to inflammation and depressive disorders. While meta-analyses suggest that PIMs are cross-sectionally associated with depression severity or MDD, and perhaps generally associated with specific symptoms such as disturbed sleep, fatigue, or alterations in appetite, prior work has raised concerns regarding the reproducibility of findings (specifically the strength and statistical significance of such relationships) [6–9]. For instance, somatic depressive symptoms (e.g., sleep, appetite, fatigue) and CRP were positively and strongly associated in the ELSA population study ($n = 5909$; $OR_{adjusted} = 1.30–1.97$) [10], but not statistically associated in the Gutenberg Health study ($n = 5000$; $OR_{adjusted} = 1.02$, $p = 0.44$) [11], while the Moli-sani study found that only one somatic symptom (appetite) was moderately associated with CRP ($n = 13,301$; $OR_{adjusted} = 1.1–1.33$) [12]. Furthermore, a major challenge in inflammatory marker research is understanding the role of confounding factors (e.g., lifestyle or risk factors for depression) influencing or contributing to both PIM elevation and specific depressive symptoms (e.g., sleep), and research findings may be heavily influenced by decisions about the number of such factors or the choice of which factors are included as covariates in analyses [7, 13]. Of note, given that PIM-depression associations may be stronger or not observed in certain subgroups, reproducible moderators of the PIM-depression relationship are also important to identify, which remains a challenge partially due to the significant variation in potential analytic decisions in such analyses [14]. Given these discrepancies and challenges, multiverse analysis could inform the extent to which results are robust to a variety of analytic choices and, conversely, what analytic options influence replicability. Identification of robust associations could guide ongoing clinical trial research into which relationships might be the most worthwhile towards ultimately targeting PIMs to improve depression symptomatology.

To address these reproducibility questions, we utilize data from the National Health and Nutrition Examination Survey (NHANES) cohort ($n = 25,962$), assessing the association between depression severity (from the Patient Health Questionnaire-9 [PHQ-9]) and two PIMs (WBC and CRP) [15]. Though associations of CRP and specific depressive symptoms have been analyzed in prior studies (representing a small number of the spectrum of all possible analytic specification strategies), few studies have investigated symptom-specific associations of WBC [12, 16]. In our multiverse analyses, we examine associations between PIMs and both total depression severity and individual depressive symptoms across eight common analytic decisions ($n = 58,320$ analytic combinations), which is the most comprehensive multiverse analysis in medicine to our knowledge. We additionally conduct exploratory multiverse analyses investigating moderators of the relationship between PIMs and depression, given the import of such relationships. We assess both effect sizes and statistical significance of associations, and robustness of results to effects of different analytic decisions. We describe how multiverse analysis can be used to identify both results that are more likely to reproduce and analytic decisions affecting the reproducibility of results.

METHODS

Participants and measures

Participants included in this analysis were 25,962 adults who provided data across any of five cycles of NHANES with valid PHQ scores and also either WBC or CRP levels, with details of this cohort in prior studies

(see Supplement S1) [14]. NHANES is a sequence of surveys administered across the United States collecting health statistics about the general population, providing cross-sectional data on participants [15].

To recruit participants (see <https://www.cdc.gov/nchs/nhanes/index.htm> for an extended explanation), NHANES, a program of studies that is part of the Centers for Disease Control, utilizes US Census information, and separates all counties in the US into 15 groups of counties with similar characteristics (e.g., based on demographic, health status, and metropolitan statistical area status variables). Then, one county is randomly selected from each group, and ~600–720 households within this county are randomly chosen. Within each household, a random number of participants within each selected household are then asked to participate in the interview. Letters are sent to eligible households or NHANES interviewers go to participants' homes to determine eligibility. Eligible individuals are then voluntarily asked to complete an interview portion (including demographic, health-related information, and self-report health measures) via telephone or in-person interview. Additionally, a subset of participants complete laboratory tests conducted separately in a local NHANES mobile examination center (MEC). The MEC is an advanced mobile clinical laboratory center, staffed by physicians, phlebotomists and health technicians, and physical examination and blood draws are conducted in the MEC. All components of NHANES surveys and transportation to the MEC are provided to participants free of charge. NHANES surveys exclude individuals in supervised care or custody in institutional settings or active-duty military personnel, but do not exclude based on age, gender, or other characteristics.

Primary measures and covariates

Depression severity was measured with the PHQ-9, a well-validated, nine-item self-report questionnaire. For total PHQ-9 scores, scores range 0 to 27, with scores >9 having strong sensitivity/specificity for a diagnosis of MDD (used to define "probably depressed" individuals in our study) [17]. In the PHQ-9, specific symptoms of depression include anhedonia, low mood, sleep disturbance, low energy, appetite disturbance, low self-esteem, concentration difficulties, psychomotor disturbances, and thoughts related to suicide/self-harm.

WBC levels were obtained through Beckman Coulter analyzers and CRP levels were obtained through Behring Nephelometer, Beckman Coulter Synchron analyzers, or Roche-Cobas 6000 analyzers (see Supplement S1 for details, including details related to CRP measures across study waves).

Secondary measures and covariates

A key analytic decision used in the subsequent multiverse analysis was related to the selection of covariates. We examined 21 covariates used in prior studies examining associations of depression and PIMs (see Supplement S2/S3 for details on covariates and covariate values across the cohort). Self-reported covariates included age, gender (male or female), marital status (married/cohabitating or not), education level (above or equivalent to/below high school), poverty index (lower scores indicate smaller income to poverty ratios), race/ethnicity (dichotomized as non-Hispanic White or not), fasting time before blood draw, any prescription drug use, antidepressant use, non-steroidal anti-inflammatory drug (NSAID) use, opiate use, statin use, total sedentary minutes, total minutes of vigorous physical activity, medical illness burden (up to 10 chronic medical illnesses; e.g., congestive heart failure), and acute illness (in the past 30 days; e.g. cold). Objectively measured covariates included body mass index (BMI), serum cotinine levels (reflecting recent nicotine exposure), HbA1C, systolic blood pressure, and diastolic blood pressure.

Multiverse decisions

Analytic choice description. To select which analytic decisions would be examined, we qualitatively reviewed the literature for analytic decisions that differed amongst published studies examining associations between PIMs (or related biological variables/factors) and total depression severity, diagnosis of depressive disorder, or individual depressive symptoms (see Table 1) [7, 13]. We included decision options identified in these studies. Each individual analysis (termed "specification") included one option from each decision. For analyses examining PIMs and PHQ total scores, we identified 58,320 potential specifications (32,400 non-overlapping specifications), with 1944 of these pertaining to linear regression models with PIM values that were not log-transformed or binned (allowing for inference testing and effect size comparison).

Table 1. Analytic decision table.

Analytic decision	Number of options	Analytic decision options	References
Covariate number	6	0, 4, 8, 12, 16, 21	[39–42]
Covariate form	3	Of 11 continuous covariates: dichotomizing 11, 6, or 0 continuous covariates based on clinically meaningful values or median value	[43, 44]
PHQ outcome/model form	5	Logistic regression (dichotomizing PHQ at 15,10, or 8), Linear regression (continuous outcome measures), or Negative binomial regression (count for the number of PHQ symptoms endorsed)	[45, 46]
PIM transformation	2	Log transform or no transformation	[47]
PIM binning	3	Quartile, tertile, or no binning	[48, 49]
Outlier treatment	6	Exclude clinical outliers (WBC, CRP, or both WBC and CRP), exclude outliers >3 standard deviations above mean, winsorize outliers, or keep outliers. Outliers identified by any method represented <15% of the data.	[7, 40, 50–52]
Gender	3	All (male and female), male, or female	[53, 54]
Exclusion of cohorts	6	Exclude based on acute illness, psychiatric medication, immune-affecting medications, any immune/psychiatric medication or illness, nondepressed, or do not exclude	[55–58]
Total	58,320		

This table describes the potential analytic decisions and the different options available for each analytic decision. Example references are provided of prior peer-reviewed published studies using such analytic decisions or review papers discussing such decisions [7, 13]. See Supplement S4.2 for further rationale for analytic decisions.

Note. Covariate selection (e.g., which covariates are selected to be included in analysis) are not included in this table because this is dependent on covariate number and with 21 covariates, ~2 million combinations of covariates exist, resulting in methodological and computational complexity beyond the scope of this analysis. However, covariate selection was examined separately through use of sampling (see Section “Effects of covariate selection on PIM-depression associations”). For covariate number, to minimize effects of the choice of covariates (given we could not explore this fully in our primary multiverse analyses due to lack of computational power), we retained a fixed order of 21 covariates (see Supplement S2).

Analytic decision rationale. Researchers might use these different decisions for a variety of reasons (see Supplement S4.D for more extensive explanations). For example, for decisions related to covariate number and covariate form (e.g., dichotomized or continuous), these might depend on the researcher’s beliefs about how influential a given a covariate is, the nature by which a covariate affects depressive symptoms (e.g., linearly related or related via a step-function), or limitations of a dataset (only certain variables/forms are easily available). As another example, for decisions related to log-transforming variables given concerns of non-normal distribution of residuals, these decisions might vary given that real-world data often does not follow normal distributions, and a multitude of methods and tests (both graphical and statistical) exist that might determine if residual distributions are normal, with variation in determination of normality depending on method used [18, 19]. Even amongst experts, discrepancy exists whether researchers should even examine normality of residuals or conduct variable transformation in cases of non-normal distribution of residuals given costs of variable transformation, such as decreased generalizability of findings related to such results [19, 20].

Analysis details

Statistical analyses were conducted using the statistical software Rv3.5.2 (see Supplement S4). We conducted multiverse analyses examining the association between total PHQ scores and PIMs and examining the association between each of nine depressive symptoms on the PHQ and PIMs. We obtained both effect sizes (beta weights for linear regression models and odds ratios for logistic regression models) and *p* values for all possible analytic specifications identified in our multiverse (see Table 1), using the PIM as the independent variable and depression severity as the dependent variable. Effect sizes and *p* values were chosen given they are the most commonly reported in the literature. Beta weights or ORs allow examination of effect size, while *p*-values allow comparison of results across different model types (e.g., logistic and linear regression). Of note, for interested readers, we also present identical multiverse analyses in probably depressed individuals (with probable depression definitions based on both PHQ total score cutoffs and DSM based criterion applied to the PHQ) in Supplement S12.

Multiverse associations of PIMs and depression. We descriptively presented median beta weights or percentage of analyses (termed “POA”) with *p* values less than 0.05 for each individual decision option [21–23].

To estimate the influence of an individual analytic choices on results as done in prior studies, we calculated Cohen’s *d* for the effect size for the difference between (a) the mean beta weight for a single decision specification and (b) the mean beta weight for all other decision specifications, descriptively noting a decision option as “influential” if $|Cohen's\ d| > 0.5$ [24–26]. To compare *p*-values between analytic specifications, we identified percentage of analyses with *p* < 0.05 for a given decision option, presenting the spread of POA between decision options for each decision.

For statistical inference, specification curve analysis was used (see Simonsohn et al. [3] and the Supplement S4.A/B/C, and S5 for extended text/graphical descriptions of analyses). This analysis involves generating all possible analytic specifications and examining if a dataset with randomly permuted PIM values (using resampling to generate 1000 shuffled samples) would be likely to have similar results using identical analytic specifications, using joint tests for inference (JTI; which utilize metrics comparing median beta weights and *p*-values) [3, 27]. Results were considered to be statistically significant by JTI if having *p* < 0.05 for both the median beta weight test (p_{MBW}) and a *p*-value based test (test based on either share of significant results [p_{SSDR}] or Stouffer’s *Z* [p_{SZ}]).

Examination of analytic decision of covariate selection. To examine decisions related to covariate selection (fully independent from prior multiverse analyses), we examined all 21 covariates in logistic regression models using participants with valid data for all 21 covariates (*n* = 15,796), using “probably depressed” status as the outcome (see Supplement S4.D for methodological details). For each “*k*” different number of covariates (i.e., 1, 4, 8, 12, 16, or 21), all possible subsets of 21 covariates using a “*k*” number of covariates (i.e., 21 choose “*k*”) were identified and up to 2000 of these combination subsets were randomly selected. ORs frequencies, median ORs, and POA for results from these combination subsets were calculated for each “*k*” number of covariates.

Examination of Individual PHQ symptoms and PIM associations. To examine associations between PIMs and individual depressive symptoms, we conducted specification curve analyses (using logistic regression models with each individual PHQ symptom score presence/absence as the outcome) and identified median ORs and POAs for each decision option. We performed both “raw” analyses (e.g., including all possible analytic decisions) and “optimized” analyses (e.g., analyses that we hypothesized

Table 2. Single decision table for beta weights and p values for WBC and CRP.

PIM	Decision	Specification	Mean B weight ^a	Influential Cohen's d (B's of specification compared to other specifications)
CRP				
	Exclusion of cohort	Exclude non-depressed	0.054	0.53
	Genders examined	Female	0.021	-0.65
	Genders examined	Male	0.05	0.61
	Number of covariates	0	0.075	1.35
	Number of covariates	16	0.019	-0.65
	Number of covariates	21	0.017	-0.65
	Outlier treatment	Exclude elevated CRP	0.056	0.62
	Outlier treatment	Exclude elevated WBC	0.019	-0.69
	Outlier treatment	Keep all	0.02	-0.66
WBC				
	Genders examined	Female	0.037	-0.77
	Genders examined	Male	0.082	0.77
	Number of covariates	0	0.119	1.66
	Number of covariates	21	0.037	-0.66
	Outlier treatment	Exclude elevated CRP	0.038	-0.63
	Outlier treatment	Keep all	0.029	-0.96

Table presents mean beta weights (average of beta weights for a single analytic specification) and influential Cohen's d (effect size of difference between beta weights of the given specification compared to all other beta weights in other specifications for the same decision), with decisions presented that retained an 'influential' Cohen's d ($|d| > 0.5$). A positive Cohen's d indicates that given analytic specification had higher beta weights compared to all other analytic specifications. For instance, for CRP and the analytic decision of gender, beta weights from analyses including only males (with a mean beta weight of 0.05) were greater than beta weights from analyses including only females or including females and males, with a Cohen's d of 0.61 for this difference. The Supplement S6 details the full single decision table.

Note: These reflect beta weights for linear regression models without modification of the PIM (e.g. no binning of PIM and no log transformation) to ensure comparability of beta weights (see Supplement S4.2 for rationale).

^aMean B weight column – This value refers to the mean B weight for all analyses that included only that given analytic specification (listed in "Specification" column).

were more theoretically accurate) for researchers who might seek to use our subset of analytic decisions. Analyses were stratified based on gender (see Supplement S4.D for rationale).

Exploratory analysis of moderators of PIM-depression associations. To exploratorily identify potential moderators of depression and PIM associations, we examined the 21 covariates previously described, and assessed if these covariates (without any dichotomization of continuous covariates) moderated associations between depression severity and PIMs (with IVs of the PIM*moderator interaction terms and DV of PHQ scores), in identical multiverse and inferential testing analyses to the primary analyses examining PIM and depression associations (see Supplement S15 for full tabular results and results by subgroup).

RESULTS

Sample characteristics

Participants in this community-based cohort ($n = 25,962$; 49.1% male; 48 ± 19 years of age; 42.5% non-Hispanic White) had mean CRP levels of 4.3 ± 8.1 mg/L, mean WBC levels of 7.3 ± 3.5 , with 8.54% of participants with PHQ scores >9 . Most covariates were significantly associated with WBC, CRP, and PHQ scores (see Supplement S3).

Multiverse analysis: PIMs and total depression severity

Summary. Across the full spectrum of analytic decisions (specifications = 1944; see Supplement S4 for details), joint tests for inference (JTI) identified that median beta weights were significantly greater in the true sample in comparison to shuffled samples ($n = 1000$ resamples) for both WBC (median $B = 0.048$, $p < 0.001$) and CRP (median $B = 0.024$, $p < 0.001$), suggesting that such associations existed independent of choice of analytic decisions. Similarly, the

true sample had a greater share of significant results in the positive direction (SSRPD) and a larger Stouffers Z for both WBC (SSRPD = 1065/1065, $p_{SSRPD} < 0.001$, $p_{SZ} < 0.001$) and CRP (SSRPD = 1053/1053, $p_{SSRPD} < 0.001$, $p_{SZ} < 0.001$).

Effect sizes. For beta weights (see Table 2), the following decision specifications were "influential" ($|d| > 0.5$) with greater beta weights for both WBC and CRP (d values for all choices are presented in the Supplement S6): analyses examining males (relative to females and males/females; $d_{CRP} = 0.61$, $d_{WBC} = 0.77$), analyses with zero covariates ($d_{CRP} = 1.35$, $d_{WBC} = 1.66$; relative to 4–21 covariates). Lower beta weights were found for analyses including all outliers ($d_{CRP} = -0.66$, $d_{WBC} = -0.96$; relative to removing all outliers based on clinical cutoffs, etc.) or analyses removing outliers based on the other PIM ($d_{CRP} = -0.69$, $d_{WBC} = -0.63$; e.g., for analyses with CRP, removing outliers based on clinical cutoffs for WBC). Such effects could compound when combining 2+ decisions (see Supplement S7). For instance, for CRP, mean beta weights in probably depressed individuals were 9 times higher in analyses excluding clinical outliers (for WBC/CRP) versus those keeping outliers.

Percentage of significant analyses (POA). We also examined percentage of analyses with $p < 0.05$ ($POA_{p < 0.05}$) (specifications = 58,320). In terms of $POA_{p < 0.05}$, 43.8% of analyses for CRP and 46.1% of analyses for WBC overall had statistically significant associations between the respective PIM and depression severity. For specific decisions across both PIMs (see Tables 1 and 3), covariate number had the greatest between-decision spread of $POA_{p < 0.05}$ (CRP spread = 77.7%, WBC spread = 55%), followed by exclusion of a given cohort (e.g., excluding participants with acute medical illness; CRP spread = 24.4%, WBC spread = 63.7%), gender

Table 3. Percentage of statistically significant analyses.

Analytic decision	Range of POA	Decision option with highest POA			Decision option with the lowest POA		
		Decision option	POA	n^a	Decision option	POA	n^a
CRP							
Number of covariates	77.7	0	90.6	11240	21	12.9	7204
PHQ form	35.9	PHQ dichotomized (>14)	62.1	8757	PHQ dichotomized (>9)	26.2	8757
Genders examined	33.1	All	56.2	13135	Female	23.1	6469
Exclusion of cohort	24.4	None (no exclusion)	55.7	12145	Exclude any illness/ medication	31.3	7914
Outlier treatment	13	Keep all	49.5	9201	Exclude elevated WBC/CRP	36.5	8002
Covariate dichotomization	4.7	Dichotomize half	45.3	8756	No modification	40.6	8756
PIM binning	1.3	Tertile	44.6	8757	Quartile	43.3	8757
PIM transformation	0.9	Log	44.2	8757	None	43.3	8757
WBC							
Exclusion of cohort	63.7	None (no exclusion)	65.6	12204	Exclude non-depressed	1.9	1018
Number of covariates	55	0	81.1	11357	21	26.2	7224
Genders examined	42.8	All	61.3	13197	Female	18.5	6502
PHQ form	23.4	PHQ continuous (NB)	54.1	8797	PHQ dichotomized (>7)	30.8	8797
Outlier treatment	21	Winsorize	54.5	9221	Exclude elevated WBC/CRP	33.5	8023
Covariate dichotomization	5.9	Dichotomize half	49.2	8797	No modification	43.3	8797
PIM binning	1.4	Quartile	46.9	8797	Tertile	45.4	8797
PIM transformation	0.6	Log	46.3	8797	None	45.8	8797

This table indicates range of percentage of analysis with $p < 0.05$ ($POA_{p < 0.05}$) across a given analytic decision, along with details of the decision option with both the lowest and highest POA. Using CRP as an example, this would indicate that (in the context of all other possible analytic decisions), analyses with 0 covariates would be statistically significant at $p < 0.05$ in 90.6% of analyses, while analyses with 21 covariates would be statistically significant at $p < 0.05$ in 12.9% of all models. Caution should be used in interpreting such differences amongst vastly different sample sizes (e.g., differing by an order of magnitude), as smaller samples might have a lower POA due to less power.

NB Negative binomial linear regression.

^aNote that mean sample size (n) is lower in some analytic decisions related to covariates due to incomplete data across the set of 21 covariates.

(CRP spread = 33.1%, WBC spread = 42.8%), PHQ form (CRP spread = 35.9%, WBC spread = 23.4%), and outlier treatment (CRP spread = 13%, WBC spread = 21%).

Effects of covariate selection on PIM-depression associations

Effects of the actual choice of specific covariate(s) on effect sizes (stratified by number of covariates) were examined by selection of “ k ” number of covariates (e.g., 1–21) using different permutations of covariates. Notable variations were seen in ranges of ORs (reflecting odds of being “probably depressed”), see Fig. 1. For instance, in 8-covariate models, ORs ranged from 1.032 to 1.191 for CRP. Practically, this means that in one model with eight covariates, for example, an individual with a CRP level that was 4.58 mg/L higher (e.g., two standard deviations of CRP) than another individual would have a 6.4% higher odds of having a probably depressed diagnosis, but in another model with eight different covariates, they might have a 38.2% higher odds of having a probably depressed diagnosis. Furthermore, within a given number of covariates, the statistical significance of certain specifications varied, such that for CRP for instance, 67.8% of 12-covariate models were statistically significant at $p < 0.05$. Importantly, across all analyses, no single covariate was particularly influential (WBC median OR spread = 1.111–1.128, CRP median OR spread = 1.062–1.077), suggesting that the choice of combination of covariates was more meaningful than the impact of individual covariates (see Supplement S9). Notably, our choice of order of covariates (in prior analyses in Sections “Sample Characteristics” and “Multiverse analysis: PIMs and total depression severity”) generally resulted in lower odds ratio in both 4- and 8-covariate models (e.g., with up to 45% lower odds ratios; see Supplement S8).

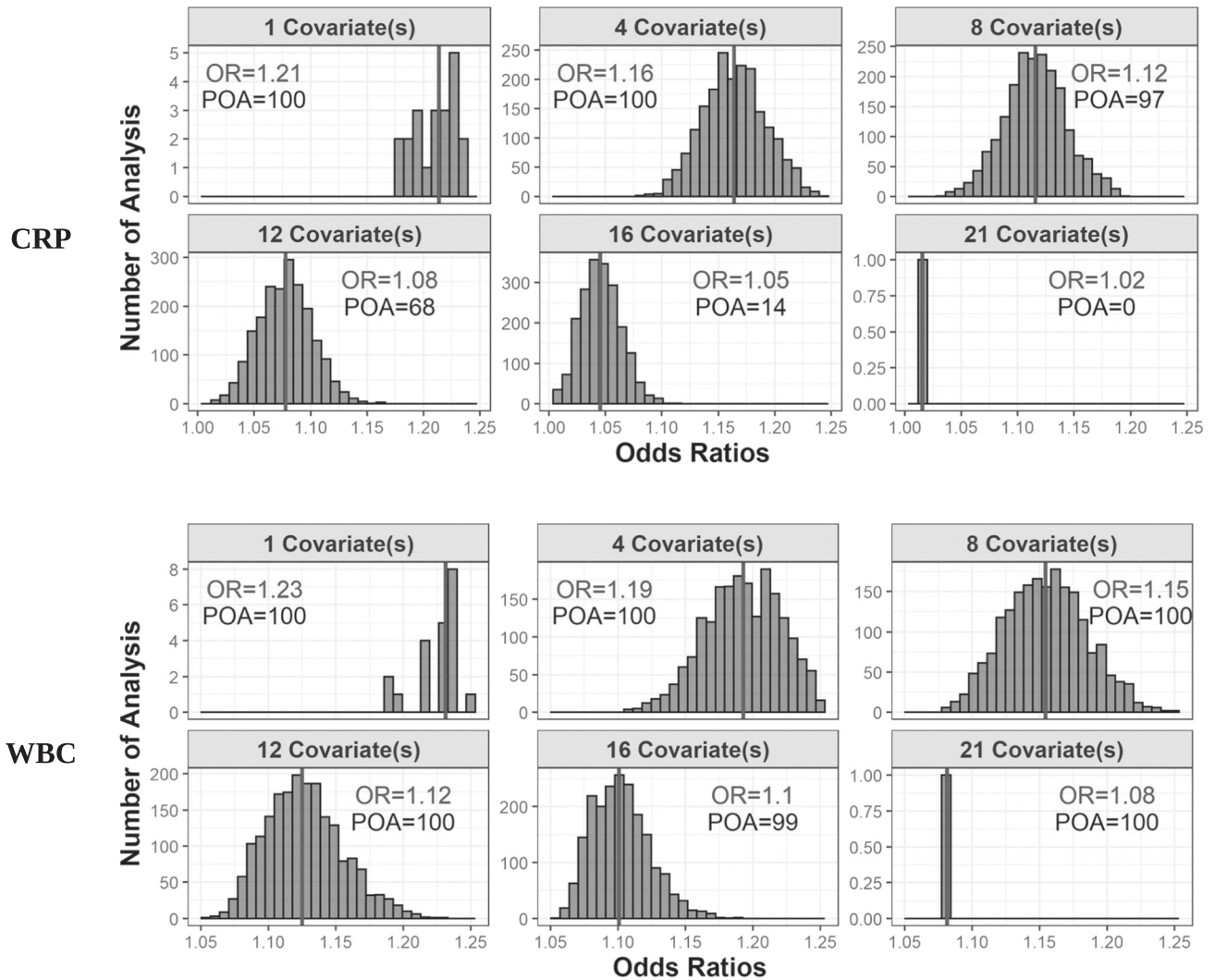
Multiverse analysis: PIMs and individual symptoms of depression

Description of analyses. Multiverse analyses were done for each individual PHQ-9 symptom stratified by gender, with presentation of both “raw” results (using all possible analytic decisions; $n = 540$ specifications; see Fig. 2) and researcher-optimized results (“opt”; $n = 18$ specifications), as an illustration of how our prior multiverse analyses could guide future analyses. In optimized analyses, we specified two analytic decisions (excluding PIM outliers based on clinical cutoffs and not excluding any groups of participants; see Supplement S4.D).

Summary of multiverse results. Based on both JTI and POAs, the most robust relationships were noted to be in males, and specifically, between WBC and anhedonia, sleep disturbance, and appetite disturbance and between CRP and sleep disturbance, low energy, and appetite disturbance ($p_{\text{joint test}} < 0.05$; POA's > 75%). Supplementary Fig. S14 emphasizes the meaningfulness of individual decisions – e.g., in examination of CRP and self-esteem associations in males, 86% of analyses were statistically significant when not excluding outliers, but 4% were statistically significant when excluding clinically-based outliers.

Multiverse symptom findings for males. For JTI in both raw and optimized analyses, for males ($n = 12,744$), positive associations between WBCs and several depressive symptoms (anhedonia, sleep disturbance, appetite disturbance, low self-esteem, psychomotor changes, and suicidality) were identified ($OR_{\text{raw}} = 1.038$ – 1.056 , $OR_{\text{opt}} = 1.042$ – 1.074) and also positive associations between CRP and low mood, sleep disturbance, low energy, poor self-esteem, and appetite disturbance ($OR_{\text{raw}} = 1.015$ – 1.024 , $OR_{\text{opt}} = 1.03$ – 1.046),

Histogram of OR Frequencies by Number of Covariates

**LEGEND**

Median ORs (blue lettering/line)

POA_{p<0.05} (black lettering)

Fig. 1 Associations of covariate choice with ORs and POA. Histograms (with number of covariates in the gray box above histogram) representing ORs frequencies across up to 2000 unique combinations of covariate for each covariate number. Text in blue (and the blue line) indicates the median ORs for each covariate set with the given number of covariates. The black text indicates POA_{p<0.05}. Thus, for example, for CRP, 100% of analyses that included different combinations of 4 covariates were statistically significant (with the median OR across all such analyses being 1.16), while 68% of analyses that included different combinations of 12 covariates were statistically significant (with the median OR across all such analyses being 1.08). *Statistical Model: Logistic regression model: PHQ probable depression (presence/absence) = $\beta_0 + \beta_1 * PIM + covariates$ (variable number).*

with Supplement S10 presenting full JTI details. For POA, in males, WBC-symptom associations with POAs > 75% (either optimized or raw) (see Supplement S11) included appetite, anhedonia, and sleep, while for CRP-symptom associations, these included energy, sleep, and appetite.

Multiverse symptom findings for females. For JTI in both raw and optimized analyses, for females ($n = 13,218$), only negative associations were noted between CRP and self-esteem/low mood

(OR_{raw} = 0.988 to 0.99, OR_{opt} = 0.972 to 0.975), with no significant associations found between WBC and any symptom. No associations had POA > 75% for females.

Exploratory multiverse analysis: moderators of the PIM and depression relationship

In multiverse analyses of moderators, the only moderator significant across inferential tests and with qualitatively moderate POAs (>25%) was race/ethnicity (for both CRP and WBC), suggesting this

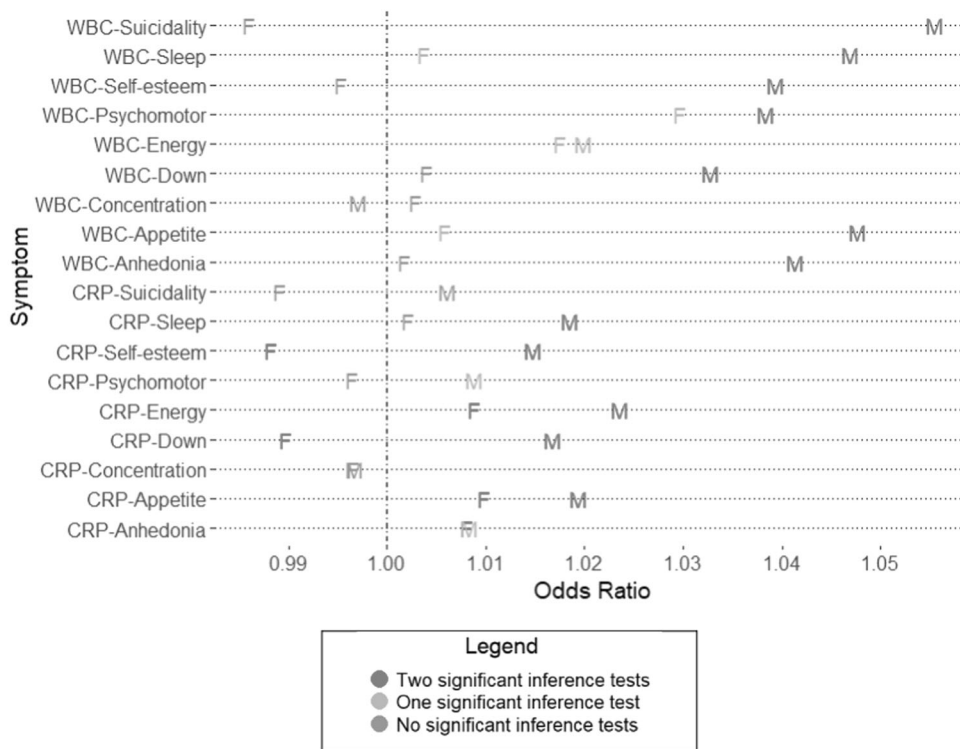


Fig. 2 Odds ratios and JTI by symptom – raw analyses. Plots of median odds ratios from specification curves amongst all participants ($n = 25,962$), reflecting “raw” unoptimized analyses. Separate ORs are provided for each gender (identified by “M” for male, and “F” for female). Thus, for example, males with a WBC level that was 1000 cells/ μL (e.g., 1 unit of WBC count) greater than the mean WBC level had a 4.7% greater odds of having sleep disturbance, while females with CRP that was 1 mg/L greater than the mean CRP level had a 1.2% lesser odds of having self-esteem concerns, with such tests significant by joint tests for inference. Note that unstandardized odds ratios are presented. **Coloring:** Coloring of letters (M/F) indicate if these specification curves were found to be significantly different from specification curves derived from randomly permuted PIM values (bootstraps = 1000), based on joint tests for inference. Red colors indicate having $p < 0.05$ on both an effect size based inference test and a p -value based inference test, yellow indicates $p < 0.05$ on either type of test (but not both), and gray indicates $p > 0.05$ for all joint tests. Thus, results with red coloring indicate that these are the most robust results against the given different analytic decisions. **Statistical Model:** Logistic regression model for each PHQ symptom: $\text{PHQ symptom presence (presence/absence)} = \beta_0 + \beta_1 * \text{PIM} + \text{covariates (variable number)}$.

might be the the most reproducible moderator (see Supplement S15). For CRP, only three moderators were robust against analytic decisions by joint tests for inference; participants with greater levels of poverty, non-Hispanic Caucasian race/ethnicity (compared to minoritized races/ethnicities), and NSAID use had stronger CRP – depression associations ($p_{\text{MBW}}'s < 0.05$, $p_{\text{SSDR}}'s < 0.05$, $p_{\text{SZ}}'s < 0.05$). For WBC, six moderators were robust against analytic decisions; participants with antidepressant use, a greater number of chronic medical illnesses, unmarried/single relationship status, greater levels of poverty, non-Hispanic Caucasian race/ethnicity (compared to minoritized races/ethnicities) and any medication use had stronger WBC – depression associations ($p_{\text{MBW}}'s < 0.05$, $p_{\text{SSDR}}'s < 0.05$, $p_{\text{SZ}}'s < 0.05$). Across all moderators, POAs ranged from 0.8 to 47.9% for WBC and 1.1 to 32.3% for CRP.

DISCUSSION

Our multiverse analysis in a large cohort ($n = 25,962$) examining associations between peripheral inflammatory markers and depression supports the overall robustness of these well-studied associations, but also lays bare the many ways in which the analytic decisions of researchers have meaningful effects on both the magnitude of effect sizes and the possibility that a given association could be statistically significant. Particularly, a wide range of analytic decisions (in our analysis, 58,000+) could be reasonably justified by equally-knowledgeable researchers for a relatively simple examination of association between PIMs and depression severity. Different choices of analytic decisions could

lead to a range of effect sizes for PIM-depression associations (e.g., up to a 10-fold increase in magnitude) and large differences in the probability that a given analysis would be statistically significant at $p < 0.05$ (e.g., ranging from 13–90%), emphasizing the effects of researcher’s degrees of freedom on the results and conclusions of studies. Although the fragility of results under certain analytic specifications is made more transparent, our findings provide insight into both which associations are more robust to different analytic decisions (e.g., more reproducible) and specific analytic decisions that are associated with greater variance in results. Together, these findings can guide future research into understanding (1) which relationships may be valuable to study more immediately (given higher reproducibility), and (2) which relationships might require more fine-grained analyses to parse the reason (e.g., methodological or scientific) for variability between analytic decisions.

Through use of joint tests for inference, we identify that CRP and WBC were associated with depressive symptoms independent of analytic decisions (for linear regression models), consistent with a vast literature for CRP and an emerging literature for WBC. However, the strength of these associations varied widely, with greater effect sizes based on gender of the cohorts (males), covariate number (fewer number of covariates), and outlier treatment (excluding or winsorizing outliers of the respective PIM). Similarly, less than 50% of all potential analyses in our well-powered dataset were statistically significant, and analyses excluding individual subgroups (e.g., based on medication use or health status), only analyzing female participants, or using a

greater number of covariates were associated with lower probabilities that a given analysis would be statistically significant [28]. Thus, researchers attempting to understand, for instance, mechanisms into the effects of an anti-inflammatory treatment on depressive symptoms, would likely have a better chance of finding a meaningful and statistically significant effect (assuming a true effect exists) in a cohort of males and using analyses which used fewer covariates and excluded PIM outliers. On the other hand, much larger sample sizes (or treatments with large effect size) might be required to detect an effect in analyses using a larger number of covariates, examining solely females, or excluding individuals based on medications/medical history. Though “ideal” analytic decisions should be purely based on scientific theory and extant evidence, the scientific evidence base frequently provides support for a variety of decisions and individual researchers’ scientific models consistently vary (see Table 1), resulting in no “ideal” set of analytic decisions and emphasizing the utility of multiverse analyses to enhance reproducibility.

Importantly, we found that the actual combination of covariates selected contributed to sizable variation in effect sizes, appearing to be independent of effects of any one individual covariate. Such results are not surprising given that cross-sectional analyses are unable to fully account for relevant complex causal, bidirectional, and/or moderation effects between PIMs, depression, and multiple covariates [14, 29]. Unfortunately, though a simplistic recommendation regarding covariate inclusion (e.g., always including covariates X, Y, and Z) would be preferable, inclusion of covariates may either be necessary (e.g., if there is true linear confounding) or invalid (e.g., if depression has causal effects on covariates), and cross-sectional analyses cannot adjudicate the “true” role of such covariates. Thus, researchers should transparently examine a wide range of subsets of potential covariates available (as done in this multiverse analysis) given this could greatly affect reproducibility of results. If researchers find variance in results based on covariate selection, then further research into such discrepancies, such as studies using experimental designs, are likely required to identify an accurate statistical model specification.

Our multiverse analyses also identified associations between PIMs and specific depressive symptoms that were robust to analytic decisions. In both raw and optimized analyses, the most reproducible relationships were in males, and specifically, between WBC and anhedonia, sleep, and appetite and between CRP and sleep, energy, and appetite. Notably, for males, reproducibility defined by effect size (based on JTI) and defined by p -values differed from one another on 5/9 symptoms, for both WBC and CRP. Though debate on the meaningfulness of effect sizes versus statistical significance based on p values exists, our multiverse analysis provides the benefit that it allows determination of reproducibility of *either* of these metrics.

Our findings in the present study are novel and impactful for several reasons. At its core, our analysis is the first multiverse analysis to examine well-studied biological markers (e.g., CRP) as they relate to depression. Hopefully, our analyses will motivate other researchers to use multiverse analyses when examining complex phenomenon such as depression. Our novel findings relate to both identifying wide variations in the strength and the presence of PIM-depression associations based on analytic decisions, and also recognition of *the degree to which these associations are reproducible* across a range of analytic decisions, by using both descriptive statistics and statistical inference. Particularly, our multiverse analyses examining PIM-symptom relationships and moderator effects allows recognition of the most reproducible relationships for these sets of analyses, which is a major challenge in the field, and which has not been done by either individual studies or meta-analyses [7]. Second, compared to other multiverse analyses (generally examining ≤ 5 decisions and ≤ 5000 analytic combinations), our analyses greatly extend the number of commonly used analytic decisions assessed (total 9

decisions), types of analytic decisions used (e.g., log-transformation, covariates), and total analytic combinations (58,320 specifications) [2, 30–32]. Third, as opposed to other individual studies and meta-analyses examining PIMs and depression, we also specifically identify *which* common analytic decisions (from a broad array of decisions) may greatly impact the strength of PIM-depression associations or statistical significance of results (e.g., genders included in analyses and number of covariates), highlighting that these decisions should be more carefully considered in future analyses to ensure reproducibility. Importantly, though larger studies such as meta-analyses can be useful, such analyses incorporate their own analytic decisions (e.g., definition of outcome, study inclusion criterion) and are limited to analytic choices and datasets present in the individual studies that comprise the meta-analysis. These factors may greatly limit the reproducibility of findings from meta-analyses, as shown by multiverse analyses of meta-analyses in other fields [33].

On an exploratory basis, we conducted multiverse analyses of different potential moderators affecting PIM and depression associations, identifying that race/ethnicity for both WBC and CRP appeared to be a reproducible moderator based on both inferential testing and POA. Several other moderators for WBC or CRP appeared to have associations that were robust across analytic decisions by joint tests for inference, including medication use, marital status, poverty level, and number of chronic illnesses. Our findings of reproducible moderators here overlapped with some moderators identified in prior studies, including research by our own group [14]. However, our results also diverged from prior findings by both identifying some reproducible moderators that were not identified in those prior isolated analyses (e.g., marital status for WBC) and also noting that previously identified moderators were not robust against analytic decisions (e.g., BMI for CRP). Our multiverse findings related to moderation analyses emphasize that any given individual analysis (including our own prior analyses!) can have fragility in results due to choices of individual analytic decisions and that multiverse analyses enhance reproducibility by making more clear which findings are not influenced by this choice of analytic decisions.

In addition to broadly identifying research findings that have greater reproducibility across different analytic decisions, many benefits of multiverse analysis exist (see Supplement S13). First, multiverse analysis can spark future research studies for both reproducible findings and less reproducible findings, exploring respectively the mechanisms and factors related to concordance or discordance. Second, multiverse analyses can guide an individual researcher’s decision-making based on identification of which decision options result in more reproducible results. Third, multiverse analyses are flexible, allowing addition and removal of decisions. Notably, even pre-registered studies, meta-analyses, randomized controlled trials, and Bayesian analyses are vulnerable to arbitrary decisions of individual researchers, potentially benefiting from multiverse analyses [34, 35]. As noted by others, these benefits of multiverse analyses will likely not be realized unless editors and peer reviewers recognize the scientific utility of the transparency of multiverse, given major disincentives against publishing multiverse analyses (e.g., presentation of results that may not be statistically significant and/or contrary to author’s hypothesis) [36].

A limitation of our analysis is that it is not exhaustive, but rather provides an initial step to identifying more reproducible relationships and increasing transparency of analyses, particularly for PIM-depression associations. We were also not able to examine all relevant decisions given the computational complexity and scope of the undertaking for a single, first-of-its-kind manuscript. Examples of other decisions include model specification choices (e.g., non-linear relationships), data imputation effects, or examining of “composite” PIM values. More broadly, these multiverse analyses do not provide researchers the “correct” analytic decision

(but rather identify how such decisions affect given associations) and the meaningfulness of analytic decisions are dependent on other analytic decisions included in the multiverse. As with any other analysis, multiverse analyses are limited by intrinsic study sample characteristics (e.g., limited variability in a measure or covariate), emphasizing the need of multiverse analyses in different study cohorts to further enhance replicability. Also, PHQ scores and some covariates were self-reported, though these variables (e.g., past medical illness) have fair concordance with clinical measures [17, 37]. Importantly, our analyses only examined cross-sectional associations between PIMs and depression, and given that a limited number of longitudinal studies have identified relationships between PIMs and prospective depression, future multiverse studies would benefit by examining these potentially meaningfully longitudinal associations [28, 38].

Other interesting findings emerged in our analyses. In our cohort including several cycles of NHANES, the rate of probable depression was 8%, yet the mean CRP level was >4 mg/L and 35% of the sample had a CRP > 3 mg/L (see Supplement S2), suggesting limited specificity of low-grade inflammation with probable depression diagnosis. Also, females had overall weaker associations between total depression severity and PIMs, as compared to males. Consistently, higher PIM levels were associated with lower levels of some specific depressive symptoms in females but higher levels of the same depressive symptoms in males. These findings emphasize the complex relationship between depression and low-grade inflammation and consequently, the need to understand, through multiverse analyses, factors that might either enhance or reduce depression-PIM associations (e.g., moderators) or how specific features of depression are related to PIMs.

Thus, in our multiverse analysis in the well-powered NHANES dataset ($n = 25,962$), we find that associations between PIMs and total depression severity exist independent of potential variation in analytic decisions. However, significant variability exists in both effect sizes of associations and probability of an individual analysis being statistically significant, based on a wide range of individual analytic decisions. At the symptom level, we also demonstrated how multiverse analysis can be used to guide and optimize analyses, identifying more reproducible PIM-symptoms associations (e.g., WBC-sleep associations in males). All study analyses include analytic decisions and thus are vulnerable to effects of such decisions, and thus, multiverse analyses improve the chance of independent replication of study results. Given that advancement in science fundamentally involves contributions by different research groups, such analyses would allow researchers to more readily build upon others' results, facilitating scientific discovery and efficiently eliminating less fruitful (e.g., less reproducible) expenditures of resources. Given the ongoing replication crisis in medicine and science, we hope that researchers, reviewers and journal editors adopt or encourage multiverse analyses, in the service of accelerating the acquisition of reliable scientific knowledge.

REFERENCES

- Baker M. Reproducibility crisis. *Nature*. 2016;533:353–66.
- Steen S, Tuerlinckx F, Gelman A, Vanpaemel W. Increasing transparency through a multiverse analysis. *Perspect Psychological Sci*. 2016;11:702–12.
- Simonsohn U, Simmons JP, Nelson LD. Specification curve: Descriptive and inferential statistics on all reasonable specifications. Available at SSRN 2694998. 2019.
- Lonsdorf T, Gerlicher A, Klingelhoefer-Jens M, Krypotos A-M. Multiverse analyses in fear conditioning research. *Behaviour Research and Therapy*. 2021;153:104072.
- Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, et al. Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. 2020;582:84–8.
- Majd M, Saunders EF, Engeland CG. Inflammation and the dimensions of depression: a review. *Front Neuroendocrinol*. 2020;56:100800.
- Horn SR, Long MM, Nelson BW, Allen NB, Fisher PA, Byrne ML. Replication and reproducibility issues in the relationship between C-reactive protein and depression: a systematic review and focused meta-analysis. *Brain Behav Immun*. 2018;73:85–114.

- Yuan N, Chen Y, Xia Y, Dai J, Liu C. Inflammation-related biomarkers in major psychiatric disorders: a cross-disorder assessment of reproducibility and specificity in 43 meta-analyses. *Transl Psychiatry*. 2019;9:1–13.
- Moriarty DP, Horn SR, Kautz MM, Haslbeck JM, Alloy LB. How handling extreme C-reactive protein (CRP) values and regularization influences CRP and depression criteria associations in network analyses. *Brain Behav Immun*. 2021;91:393–403.
- White J, Kivimäki M, Jokela M, Batty GD. Association of inflammation with specific symptoms of depression in a general population of older people: the English Longitudinal Study of Ageing. *Brain Behav Immun*. 2017;61:27–30.
- Michal M, Wiltink J, Kirschner Y, Wild PS, Münzel T, Ojeda FM, et al. Differential associations of depressive symptom dimensions with cardio-vascular disease in the community: results from the Gutenberg health study. *PLoS One*. 2013;8:e72014.
- Gialluisi A, Di Castelnuovo A, Bracone F, De Curtis A, Cerletti C, Donati MB, et al. Associations between systemic inflammation and somatic depressive symptoms: findings from the Moli-sani study. *Depress Anxiety*. 2020;37:935–43.
- O'Connor M-F, Bower JE, Cho HJ, Creswell JD, Dimitrov S, Hamby ME, et al. To assess, to control, to exclude: effects of biobehavioral factors on circulating inflammatory markers. *Brain Behav Immun*. 2009;23:887–97.
- Rengasamy M, Da Costa E, Silva SA, Spada M, Price RB. Does the moderator matter? Identification of multiple moderators of the association between peripheral inflammatory markers and depression severity in a large racially diverse community cohort. *Neuropsychopharmacology*. 2022;47:1693–701.
- NHANES, National Center for Health Statistics. National Health and Nutrition Examination Survey. 2021. Centers for Disease Control and Prevention website, NHANES, National Center for Health Statistics; 2021. <https://www.cdc.gov/nchs/nhanes/index.htm>.
- Milaneschi Y, Kappelmann N, Ye Z, Lamers F, Moser S, Jones PB, et al. Association of inflammation with depression and anxiety: evidence for symptom-specificity and potential causality from UK Biobank and NESDA cohorts. *Molecular Psychiatry*. 2021;26:7393–402.
- Moriarty AS, Gilbody S, McMillan D, Manea L. Screening and case finding for major depressive disorder using the Patient Health Questionnaire (PHQ-9): a meta-analysis. *Gen Hospital Psychiatry*. 2015;37:567–76.
- Riniolo TC, Porges SW. Evaluating group distributional characteristics: Why psychophysiology should be interested in qualitative departures from the normal distribution. *Psychophysiology*. 2000;37:21–8.
- Knief U, Forstmeier W. Violating the normality assumption may be the lesser of two evils. *Behav Res Methods*. 2021;53:2576–90.
- Changyong F, Hongyue W, Najji L, Tian C, Hua H, Ying L. Log-transformation and its implications for data analysis. *Shanghai Arch Psychiatry*. 2014;26:105.
- Wright L, Head JA, Jivraj S. How robust is the association between youth unemployment and later mental health? An analysis of longitudinal data from English schoolchildren. *Occup Environ Med*. 2021;78:618–20.
- Harder JA. The multiverse of methods: extending the multiverse analysis to address data-collection decisions. *Perspect Psychological Sci*. 2020;15:1158–77.
- Moore MJ, Demeyere N. Multiverse to inform neurological research: an example using recovery outcome of neglect. *J Neurol*. 2022;269:233–42.
- Schäfer T, Schwarz MA. The meaningfulness of effect sizes in psychological research: differences between sub-disciplines and the impact of potential biases. *Front Psychol*. 2019;10:813.
- Szucs D, Ioannidis JP. Empirical assessment of published effect sizes and power in the recent cognitive neuroscience and psychology literature. *PLoS Biol*. 2017;15:e2000797.
- Modecki KL, Low-Choy S, Uink BN, Vernon L, Correia H, Andrews K. Tuning into the real effect of smartphone use on parenting: a multiverse analysis. *J Child Psychol Psychiatry*. 2020;61:855–65.
- Barendse ME, Byrne ML, Flournoy JC, McNeilly EA, Guazzelli Williamson V, Barrett A-MY, et al. Multimethod assessment of pubertal timing and associations with internalizing psychopathology in early adolescent girls. *J Abnormal Psychol*. 2021;131:14.
- Mac Giollabhui N, Ng TH, Ellum LM, Alloy LB. The longitudinal associations of inflammatory biomarkers and depression revisited: systematic review, meta-analysis, and meta-regression. *Mol Psychiatry*. 2020;26:3302–14.
- Schisterman EF, Cole SR, Platt RW. Overadjustment bias and unnecessary adjustment in epidemiologic studies. *Epidemiology*. 2009;20:488.
- Lewis MW, et al. Multiverse analyses of fear acquisition and extinction retention in posttraumatic stress disorder. *Psychophysiology*. 2023:e14265.
- Bloom PA, VanTieghem M, Gabard-Durnam L, Gee DG, Flannery J, Caldera C, et al. Age-related change in task-evoked amygdala—prefrontal circuitry: a multiverse approach with an accelerated longitudinal cohort aged 4–22 years. *Hum Brain Mapp*. 2022;43:3221–44.
- Black L, Panayiotou M, Humphrey N. Internalizing symptoms, well-being, and correlates in adolescence: a multiverse exploration via cross-lagged panel network models. *Dev Psychopathol*. 2022;34:1477–91.

33. El Bahri M, Wang X, Biaggi T, Falissard B, Naudet F, Barry C. A multiverse analysis of meta-analyses assessing acupuncture efficacy for smoking cessation evidenced vibration of effects. *J Clin Epidemiol.* 2022;152:140–50.
34. Olsson-Collentine A, van Aert R, Bakker M, Wicherts J. Meta-analyzing the multiverse: A peek under the hood of selective reporting. *Psychol Methods.* 2023. <https://psycnet.apa.org/fulltext/2023-71132-001.html>.
35. Simmons JP, Nelson LD, Simonsohn U. False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Sci.* 2011;22:1359–66.
36. Fanelli D. Do pressures to publish increase scientists' bias? An empirical support from US States Data. *PLoS One.* 2010;5:e10271.
37. Okura Y, Urban LH, Mahoney DW, Jacobsen SJ, Rodeheffer RJ. Agreement between self-report questionnaires and medical record data was substantial for diabetes, hypertension, myocardial infarction and stroke but not for heart failure. *J Clin Epidemiol.* 2004;57:1096–103.
38. Moriarty DP, Kautz MM, Mac Giollabhui N, Klugman J, Coe CL, Ellman LM, et al. Bidirectional associations between inflammatory biomarkers and depressive symptoms in adolescents: Potential causal relationships. *Clin Psychological Sci.* 2020;8:690–703.
39. Xue Y, Liu G, Geng Q. Associations of cardiovascular disease and depression with memory related disease: a Chinese national prospective cohort study. *J Affect Disord.* 2020;266:187–93.
40. Bondy E, Norton SA, Voss M, Marks RB, Boudreaux MJ, Treadway MT, et al. Inflammation is associated with future depressive symptoms among older adults. *Brain Behav Immun Health.* 2021;13:100226.
41. Manfro PH, Anselmi L, Barros F, Gonçalves H, Murray J, Oliveira IO, et al. Youth depression and inflammation: Cross-sectional network analyses of C-Reactive protein, interleukin-6 and symptoms in a population-based sample. *J Psychiatric Res.* 2022;150:197–201.
42. Frank P, Jokela M, Batty GD, Cadar D, Steptoe A, Kivimäki M. Association Between Systemic Inflammation and Individual Symptoms of Depression: A Pooled Analysis of 15 Population-Based Cohort Studies. *Am J Psychiatry.* 2021;178:1107–18.
43. Pitharouli MC, Hagenaars SP, Glanville KP, Coleman JR, Hotopf M, Lewis CM, et al. Elevated C-reactive protein in patients with depression, independent of genetic, health, and psychosocial factors: results from the UK Biobank. *Am J Psychiatry.* 2021;178:522–29.
44. Lee S, Oh SS, Jang S-I, Park E-C. Sex difference in the association between high-sensitivity C-reactive protein and depression: the 2016 Korea National Health and Nutrition Examination Survey. *Sci Rep.* 2019;9:1–10.
45. Manea L, Gilbody S, McMillan D. Optimal cut-off score for diagnosing depression with the Patient Health Questionnaire (PHQ-9): a meta-analysis. *CMAJ.* 2012;184:E191–6.
46. Tracy M, Morgenstern H, Zivin K, Aiello AE, Galea S. Traumatic event exposure and depression severity over time: results from a prospective cohort study in an urban area. *Soc Psychiatry Psychiatr Epidemiol.* 2014;49:1769–82.
47. Khandaker GM, Zuber V, Rees J, Carvalho L, Mason AM, Foley CN, et al. Shared mechanisms between coronary heart disease and depression: findings from a large UK general population-based cohort. *Mol Psychiatry.* 2020;25:1477–86.
48. Linkas J, Ahmed LA, Csifcsak G, Emaus N, Furberg A-S, Grimnes G, et al. Are pro-inflammatory markers associated with psychological distress in a cross-sectional study of healthy adolescents 15–17 years of age? The Fit Futures study. *BMC Psychol.* 2022;10:1–13.
49. Cong X, Tracy M, Edmunds LS, Hosler AS, Appleton AA. The relationship between inflammatory dietary pattern in childhood and depression in early adulthood. *Brain Behav Immun Health.* 2020;2:100017.
50. Burrows K, Stewart JL, Kuplicki R, Figueroa-Hall L, Spechler PA, Zheng H, et al. Elevated peripheral inflammation is associated with attenuated striatal reward anticipation in major depressive disorder. *Brain Behav Immun.* 2021;93:214–25.
51. Stewart JC, Rand KL, Muldoon MF, Kamarck TW. A prospective evaluation of the directionality of the depression–inflammation relationship. *Brain Behav Immun.* 2009;23:936–44.
52. Raison CL, Rutherford RE, Woolwine BJ, Shuo C, Schettler P, Drake DF, et al. A randomized controlled trial of the tumor necrosis factor antagonist infliximab for treatment-resistant depression: the role of baseline inflammatory biomarkers. *JAMA Psychiatry.* 2013;70:31–41.
53. Lamers F, Milaneschi Y, Smit JH, Schoevers RA, Wittenberg G, Penninx BW. Longitudinal Association Between Depression and Inflammatory Markers: Results From the Netherlands Study of Depression and Anxiety. *Biol Psychiatry.* 2019;85:829–37.
54. Tayefi M, Shafiee M, Kazemi-Bajestani SMR, Esmaeili H, Darroudi S, Khakpour S, et al. Depression and anxiety both associate with serum level of hs-CRP: a gender-stratified analysis in a population-based study. *Psychoneuroendocrinology.* 2017;81:63–9.
55. Bai Y-M, Chiou W-F, Su T-P, Li C-T, Chen M-H. Pro-inflammatory cytokine associated with somatic and pain symptoms in depression. *J Affect Disord.* 2014;155:28–34.
56. Schmidt FM, Schröder T, Kirkby KC, Sander C, Suslow T, Holdt LM, et al. Pro-and anti-inflammatory cytokines, but not CRP, are inversely correlated with severity and symptoms of major depression. *Psychiatry Res.* 2016;239:85–91.
57. Birur B, Amrock EM, Shelton RC, Li L. Sex differences in the peripheral immune system in patients with depression. *Front Psychiatry.* 2017;8:108.
58. Köhler-Forsberg O, Buttenschøn HN, Tansey KE, Maier W, Hauser J, Dernovsek MZ, et al. Association between C-reactive protein (CRP) with depression symptom severity and specific depressive symptoms in major depression. *Brain Behav Immun.* 2017;62:344–50.

AUTHOR CONTRIBUTIONS

Author MR contributed to all aspects of the manuscript, including conceptualization, study design, data curation, data analysis, data interpretation and writing. Authors DM, TK, BT, and RP contributed to conceptualization, data interpretation and writing.

FUNDING

This research was supported by funding from the Ruth L. Kirschstein National Research Service Award Institutional Research Training Grants sponsored by the National Institutes of Health (Grant No. NIH T32 MH018951: Brent).

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41386-023-01621-4>.

Correspondence and requests for materials should be addressed to Manivel Rengasamy.

Reprints and permission information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Springer Nature or its licensor (e.g. a society or other partner) holds exclusive rights to this article under a publishing agreement with the author(s) or other rightsholder(s); author self-archiving of the accepted manuscript version of this article is solely governed by the terms of such publishing agreement and applicable law.