Supplementary Information

Title

Heightened amygdala reactivity and increased stress generation predict internalizing symptoms in adults following childhood maltreatment

Authors & Affiliations

[a, b] Mattia I Gerin, M.Sc., [a] Essi Viding, Ph.D., [a,c,d] Jean-Baptiste Pingault, Ph.D., [a, b] Vanessa B Puetz, Ph.D., [e] Annchen R. Knodt, M.S., [e] Spenser R. Radtke, M.S.W., [e] Bartholomew D. Brigidi, Ph.D., [f] Johnna R. Swartz, Ph.D., [e] Ahmad R. Hariri, Ph.D., [a, b] Eamon J McCrory, Ph.D.

[a] Department of Clinical Educational and Health Psychology, Division of Psychology and Language Science, University College London, London, UK; [b] Anna Freud National Centre for Children and Families, London, UK; [c] Social Genetic and Developmental Psychiatry Centre, King's College London, London, UK; [d] CESP, Univ. Paris-Sud, UVSQ, INSERM, Université Paris-Saclay, Villejuif, France; [e] Laboratory of NeuroGenetics, Department of Psychology & Neuroscience, Duke University, Durham, NC, USA; [f] Department of Human Ecology, University of California at Davis, Davis, CA, USA.

Supplementary Methods

Participants

DNS Dataset

Participants were young adult college students who were recruited as part of the ongoing Duke Neurogenetics Study (DNS; n=1144). Exclusion criteria included i) medical diagnoses of stroke, diabetes, cancer, chronic kidney or liver disease; ii) use of psychotropic, glucocorticoid, or hypolipidemic medication; iii) lifetime history of psychotic symptoms iv) conditions affecting cerebral blood flow and metabolism (e.g., hypertension); and v) did not meet quality control criteria for functional MRI scanning (see "fMRI analysis" section below for more details).

Participants with Longitudinal Data

Only individuals who had post-baseline assessment of internalizing symptoms and stressful life events were considered for this study (n=584). Note that previous studies [1] that used a largely overlapping DNS dataset and variables of interests showed that participants who completed follow-up assessment did not differ from participants who did not in relation to several variables of interest and covariates (e.g. age, childhood trauma, baseline measures of stressful life events, amygdala activity and internalizing symptoms). Participants were contacted every 3 months by email to voluntarily fill in a checklist about their current internalizing symptoms and experience of stressful life events since the previous assessment. Because the DNS is an ongoing research project, there is great variability among participants in terms of both the number of post-baseline assessment completed and, as a result, also in the time-interval lapsed between baseline and last post-baseline assessment. Therefore, only the first three post-baseline assessments (where available) for each participant were included in order to reduce variability in both the number of assessments and time elapsed between baseline and last follow-up.

Maltreated Group (MT)

In line with studies of childhood maltreatment prevalence [2,3], among the 584 participants with longitudinal available data, a subset of individuals (n=100) reported experiences of significant childhood abuse and/or neglect. This was operationalized as having experienced at least one form of childhood abuse or neglect - i.e. they scored, as established by the Childhood Trauma Questionnaire (CTQ) manual, within or above the "Moderate-Severe" range in one or more maltreatment subtype scales (i.e. emotional neglect ≥ 13, physical abuse ≥ 10, sexual abuse ≥ 8, emotional neglect ≥ 15 and physical neglect ≥ 10). In the MT group 33% experienced emotional abuse, 29% experiences physical abuse, 19% sexual abuse, 32% emotional neglect and 44% physical neglect. Moreover, 41% had experiences more than one maltreatment type. In relation to normative epidemiological data thresholds, all individuals in the MT group had at least one maltreatment subtype CTQ scores above the 90th percentile [4].

Non-Maltreated Group (Non-MT)

Participants were included in the Non-Maltreated (Non-MT; n=127) group if i) their CTQ total score was below the 50th normative percentile threshold [4] AND ii) if they scored in each CTQ maltreatment subtype scale within the none-or-minimal range [5]. 353 individuals were excluded from further analyses because they reported some experience of childhood maltreatment - thus, could not be included in the Non-MT group – nor did they meet threshold for significant experiences of abuse or neglect.

Final Sample after Propensity Score Matching

After Propensity Score Matching (PSM, described in detail below) and outlier removal, all analyses were performed using the Maltreated Group (MT; n=100) and the propensity score-matched Control Group (CT; n=96) - i.e. the Non-MT group *after* PSM.

Propensity Score Matching (PSM)

The impact of potentially confounding variables can be reduced using Propensity Score Matching (PSM). This comprises a range of statistical approaches that can be applied, prior to any

inferential statistical analysis, in order to balance the distribution of covariates across the treatment and control groups [6]. When applying PSM, several matching methodologies should be explored to identify which one yields the greatest reduction in distance (i.e. distribution of covariates) between the control and treatment groups [7–9]. The most common matching procedures rely on exact pairing, weighting or sub-classification (or a combination of those). PSM can be used with a large number of control variables without incurring in model over-fitting issues and multicollinearity, which, in traditional linear regression modelling, can reduce the ability to assess the impact of a predictor variable on the outcome variables [7,9]. Crucially for this study, unlike traditional co-varying methods in linear regression, PSM can be used reliably also for variables that are characterized by considerable distribution differences across groups [10,11].

The outcome of the matching (i.e. the balance between the two groups post-matching), is often assessed using a standardized mean difference (i.e. effect size) of the propensity score [7,8]. Although there is not a consensus on the cut-off, it has been suggested that value between 0.1 and 0.25 represent acceptable cut-offs for standardized means differences post-matching [12]. It has also been suggested that standardized mean differences for each variable pre- and post-matching can be interpreted as effect sizes, with values smaller than 0.2 considered a small difference, 0.4 medium and 0.8 large [6,10].

The R software package MatchIT [7] was used to implement four PSM methods that use different algorithms to match participants. Nearest Matching and Optimal Matching are similar procedure as they use 1:1 matching approach. In Nearest Matching, for each participant in the maltreated group one or more participants with the closest propensity score (i.e. smallest distance) is selected from the non-maltreated group. In Optimal Matching the 1 to 1 matching is conducted in such a way to minimize the overall distance between the two groups. Genetic Matching and Full Matching, which enables flexible matching within subclasses by applying weighting, were also explored.

Within the current study, the following potentially confounding variables were selected for the PSM between the MT and Non-MT groups: age, gender, IQ, ethnicity, socio-economic status (measured by parental level of education), and baseline internalizing symptoms - which was assessed by the composite score on the Mood and Anxiety Symptoms Questionnaire (MASQ) [13].

Baseline Neuroimaging Procedure and Data Analysis

Task

The fMRI paradigm used here has been shown to robustly elicit amygdala responses over a wide variety of protocols and populations [14–19]. The paradigm consisted of four task blocks interleaved with five control blocks. A total of four emotion categories were used for each task block: fearful (F), angry (A), surprised (S), and neutral (N), taken from a standardized facial expression set [20]. Participants viewed the task blocks in one of four randomly assigned orders as determined by a Latin Square (i.e., FNAS, NFSA, ASFN, SANF). During task blocks, participants viewed a trio of faces and matched one of two faces identical to a target face. Each trial in the task blocks lasted for 4 seconds with a variable interstimulus interval of 2-6 seconds (mean = 4 seconds), for a total block length of 48 seconds. The experimental blocks were interleaved with sensorimotor control shape-matching blocks, during which six geometric shape trios were presented for 4 seconds with a fixed interstimulus interval of 2 seconds for a total block length of 36 seconds. Each block was preceded by a brief instruction ("Match faces" or "Match shapes") lasting 2 seconds resulting in a total task time of 390 seconds. In the present study, we restricted analyses to fearful and angry expressions.

fMRI acquisition

Participants were scanned using one of two identical General Electric MR570 3T scanners at the Duke-UNC Brain Imaging and Analysis Centre equipped with high power high duty cycle 50mT/m gradients at 200 T/m/s slew rate, and an 8 channel head coil for parallel imaging at high bandwidth up to 1 MHz. Thirty-four interleaved AC-PC aligned axial functional slices were acquired for full-brain coverage using an inverse-spiral pulse sequence to reduce susceptibility artefacts (TR = 2000ms; TE = 30 ms; flip angle = 60°; FOV = 240 mm; 3.75 × 3.75 × 4 mm voxels; interslice skip = 0). Four initial volumes were acquired and discarded to achieve steady state equilibrium. A semi-automated high-order shimming program was used to ensure global field homogeneity. High resolution 3D structural images (TR = 7.7 s; TE; 3.0 ms; flip angle = 12°; voxel size = 0.9 × 0.9 × 4 mm; FOV = 240 mm, interslice skip = 0) were also acquired in 34 axial slices to assist registration of functional data.

fMRI analysis

Neuroimaging analyses were conducted using the software package SPM8 (www.fil.ion.ucl.ac.uk/spm/software/ spm8) implemented in Matlab 2015a (MathWorks Inc.). A standard pre-processing procedure was implemented in line with previously published research from the Duke Neurogenetics Study [e.g. 37,52]. This included head motion corrections (i.e. realigning the images to the first volume in the time series); spatial normalization into standard stereotactic space (Montreal Neurological Institute (MNI) template) using a 12-parameter affine model (final resolution of functional images = 2 mm isotropic voxels); and smoothing with a full width and half maximum 6mm Gaussian filter. Voxel-wise intensities were ratio normalized to the whole-brain global mean signal.

An artifact detection software (ART - http://www.nitrc.org/projects/artifact_detect) was implemented to create a regressor which assigned lower weighting to i) individuals volumes where scan-to-scan movement was greater than 2^0 rotation or 2mm translation, ii) volumes exhibiting significant variation in mean-volume signal intensity (i.e. volumes with mean signal smaller or greater than four standard deviations of the mean signal of all the volumes in the time series). Moreover, only data from participants with the following characteristics were included: i) ≥ 90% coverage of signal within the anatomically-defined bilateral amygdala region of interest, ii) <5% volumes exceed Artifact detection criteria for motion or signal intensity outliers, and iii) accuracy ≥75% on the matching task performed during scanning.

In line with procedures implemented in prior published data from the DNS [1,21], parameter estimates for each participant were extracted from the functional cluster (i.e. set of contiguous voxels activated at P<.05, FWE corrected – the minimum number of contiguous voxels was 10) within the anatomical amygdala (defined structurally by the Automated Anatomical Labeling Atlas -AAL).

Outlier Removal

In order to remove outliers for normally distributed outcome variables (i.e. baseline amygdala reactivity) Tukey's box-plot Inter-quartile range (IQR) method was implemented, using a multiplier of 2.2, as suggested by simulation estimates by Hoaglin & Iglewixcz (1987) [22]. For non-normally distributed data (i.e. prospective stressful life events and longitudinal internalizing symptoms levels), the 'adjusted boxplot' method developed by Hubert & Vandervieren (2008) [23] was implemented instead [24]. Consequently 12 participants were removed from further analyses (4 belonging to the MT group and 8 to the CT group).

Supplementary Results

Propensity Score Matching (PSM) Output

Optimal and Nearest Neighbor Matching did not yield satisfactory results as, after matching, the standardized mean difference between the two groups remained almost unchanged (0.32). Moreover, the absolute standardized mean difference after matching was still moderate-to-large for several covariates, such as baseline internalizing symptoms (i.e. > 0.7). Genetic Matching also did not generate satisfactory results because, despite achieving low overall distance reduction between the two groups, it discarded several individuals in the Non-Maltreated (Non-MT) group (ratio of 1 : 0.4).

However, Full Matching achieved satisfactory results. Reduction in the overall standardized mean difference across all covariates was achieved (from 0.42 to 0.09 post-matching - Table 1); moreover, the standardized mean difference for each variable was below the small effect size threshold (i.e. < 0.2), with absolute values ranging between 0.02 and 0.15. Notably, the standardized mean difference for internalizing symptoms and ethnicity (i.e. the two variables with the largest absolute standardized mean difference before matching) decreased from, respectively, 1.02 to 0.02 and from 0.70 to 0.12. Furthermore, gender and ethnicity, which are the two variables with the largest absolute standardized mean difference post-matching (0.15 and 0.11 respectively), showed a small and non-significant group difference: the CT group had 7% more females ($X^2(1) = 1.45$, $p = .29$) and 5.5% more Caucasians than the MT group ($X^2(1) = 0.57$, $p = .45$) (Table S1).

Participants characteristics for matched variables

In addition to inspecting standardized mean difference *after* PSM, we further tested whether any significant differences remained between the two groups. T-test and chi-square tests revealed that across all propensity score matched variables there were not significant differences between the MT and CT groups (Table S1).

Supplementary Table S1.
Mean and standard deviation for propensity score matched variables post-matching and p-values

	Mean (Standard Deviation) or Percentage		*p*
	MT	CT (Non-MT *Post* Matching)	
Age	19.41 (2.33)	19.57 (1.11)	.53
Gender (female)	63 %	70%	.29
IQ	119.63 (14.69)	120.47 (8.2)	.63
Ethnicity (Caucasian)	30 %	35.5 %	.45
SES	7.53 (1.8)	7.36 (1.82)	.50
Baseline Internalising symptoms	127.65 (33.02)	127.07 (36.19)	.91

Participants characteristics for non-matched variables

This subsection reports variables which were not propensity score matched (i.e. CTQ scores and longitudinal assessments time-interval).

Longitudinal time-interval information

The time-lapsed between baseline and last post-baseline assessment between the MT group (mean = 11.8 months, s.d. = 7.7, min = 1.2, max = 43.9) and the CT group (mean = 11.4 months, s.d. = 7.9, min = 2.7, max = 45.7) was not significantly different $t(197) = -0.37$, $p = .71$. Moreover, a similar number of individuals – about half - in both the MT (47.5%) and CT (50.6%) had one longitudinal assessment, while the other half of participants in each group had 2 or 3 longitudinal assessments.

CTQ scores

The MT and the Non-MT individuals, as discussed in the methods section, were selected based on their CTQ scores. Thus, the CTQ total and subscale scores for the MT and CT group (i.e. the Non-MT group post-matching) were statistically significant (Table S2). Notably, while the CTQ total

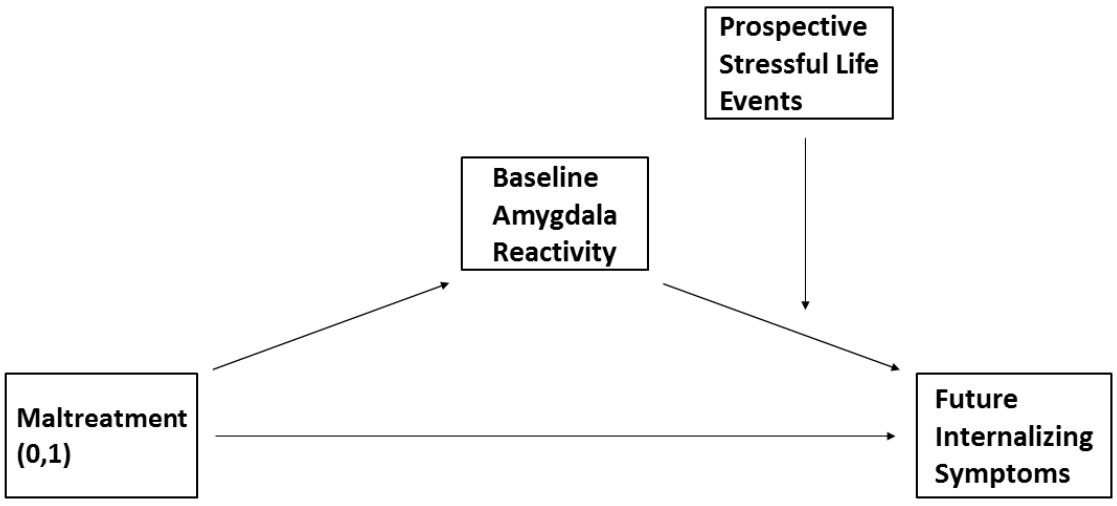
score for the MT group was above the 90th percentile of the population normative threshold, the CT group total score was close to the 25th lowest percentile of the population normative threshold [4].

Supplementary Table S2
Mean and standard deviation for CTQ values post-matching

	Mean (Standard Deviation)		*p*
	MT	CT (Non-MT *Post* Matching)	
CTQ Total Score	46.23 (8.58)	25.56 (0.5)	<.001
Physical Abuse	7.97 (3.43)	5.03 (.17)	<.001
Emotional Abuse	10.60 (4.05)	5.19 (.39)	<.001
Sexual Abuse	6.24 (2.78)	5 (0)	<.001
Physical Neglect	9.07 (3.15)	5 (0.03)	<.001
Emotional Neglect	12.35 (4.05)	5.34 (0.48)	<.001

Moderated Mediation

The moderated mediation analysis was performed in order to investigate whether the mediating effect of baseline amygdala reactivity on the relationship between maltreatment status and future internalizing symptoms was conditional on post-baseline exposure to stressful life events. In particular, as depicted in Figure S1 below, we tested whether the pathway from baseline amygdala threat reactivity (the mediator) to future internalizing symptoms was modulated by stressful life events - for further details on this model see Preacher, Rucker & Hayes (2007) model 3 [25].



Supplementary Figure S1. Conceptual moderated mediation model depicting the putative moderating

effect of post-baseline stressful life events on the mediating effect of baseline amygdala threat reactivity on the association between maltreatment status and future internalizing symptoms.

Supplementary References

1. Swartz JR, Knodt AR, Radtke SR, Hariri AR. A neural biomarker of psychological vulnerability to future life stress. *Neuron*. 2015;85(3):505-511. doi:10.1016/j.neuron.2014.12.055.

2. Radford L, Corral S, Bradley C, et al. *Child Abuse and Neglect in the UK Today: Research into the Prevalence of Child Maltreatment in the United Kingdom*. London; 2011.

3. Stoltenborgh M, Bakermans-Kranenburg MJ, Alink LRA, van IJzendoorn MH. The Prevalence of Child Maltreatment across the Globe: Review of a Series of Meta-Analyses. *Child Abus Rev*. 2015;24(1):37-50. doi:10.1002/car.2353.

4. Scher CD, Stein MB, Asmundson GJG, Mccreary DR, Forde DR. The Childhood Trauma Questionnaire in a community sample: Psychometric properties and normative data. *J Trauma Stress*. 2001;14(4):843-857. doi:10.1023/A:1013058625719.

5. Bernstein DP, Fink L. *Childhood Trauma Questionnaire: A Retrospective Self-Report*. San Antonio, TX: The Psychological Corporation; 1997.

6. McCaffrey DF, Ridgeway G, Morral AR. Propensity Score Estimation With Boosted Regression for Evaluating Causal Effects in Observational Studies. *Psychol Methods*. 2004;9(4):403-425. doi:10.1037/1082-989X.9.4.403.

7. Ho DE, Imai K, King G, Stuart EA. MatchIt : Nonparametric Preprocessing for Parametric Causal Inference. *J Stat Softw*. 2011;42(8):1-28. doi:10.18637/jss.v042.i08.

8. Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Polit Anal*. 2007;15(3):199-236. doi:10.1093/pan/mpl013.

9. Stuart EA. Matching methods for causal inference: A review and a look forward. *Stat Sci a Rev J Inst Math Stat*. 2010;25(1):1-21. doi:10.1214/09-STS313.

10. Pingault JB, Cote SM, Petitclerc A, Vitaro F, Tremblay RE. Assessing the independent contribution of maternal educational expectations to children's educational attainment in early adulthood: A propensity score matching analysis. *PLoS One*. 2015;10(3). doi:10.1371/journal.pone.0119638.

11. Rubin DC. Using prpensity scores to help design ovservational studies: Application to the tobacco litigation. *Health Serv Outcomes Res Methodol*. 2002;2:169-188. doi:10.1023/A:1020363010465.

12. Stuart EA, Lee BK, Leacy FP. Prognostic score-based balance measures can be a useful diagnostic for propensity score methods in comparative effectiveness research. *J Clin Epidemiol*. 2013;66(8 SUPPL.8). doi:10.1016/j.jclinepi.2013.01.013.

13. Watson D, Clark LA, Weber K, Assenheimer JS, Strauss ME, McCormick RA. Testing a tripartite model: II. Exploring the symptom structure of anxiety and depression in student, adult, and patient samples. *J Abnorm Psychol*. 1995;104(1):3-14. doi:10.1037/0021-843X.104.1.3.

14. Hariri AR, Mattay VS, Tessitore A, et al. Serotonin transporter genetic variation and the response of the human amygdala. *Science*. 2002;297(5580):400-403. doi:10.1126/science.1071829.

15. Hariri AR, Drabant EM, Munoz KE, et al. A susceptibility gene for affective disorders and the response of the human amygdala. *Arch Gen Psychiatry*. 2005;62(2):146-152. doi:10.1001/archpsyc.62.2.146.

16. Fisher PM, Meltzer CC, Ziolko SK, et al. Capacity for 5-HT1A-mediated autoregulation predicts amygdala reactivity. *Nat Neurosci*. 2006;9(11):1362-1363. doi:10.1038/nn1780.

17. Fisher PM, Meltzer CC, Price JC, et al. Medial prefrontal cortex 5-HT(2A) density is correlated with amygdala reactivity, response habituation, and functional coupling. *Cereb Cortex*. 2009;19(11):2499-2507. doi:10.1093/cercor/bhp022.

18. Manuck SB, Brown SM, Forbes EE, Hariri AR. Temporal stability of individual differences in amygdala reactivity. *Am J Psychiatry*. 2007;164(10):1613-1614. doi:10.1176/appi.ajp.2007.07040609.

19. Zhou Z, Zhu G, Hariri AR, et al. Genetic variation in human NPY expression affects stress response and emotion. *Nature*. 2008;452(7190):997-1001. doi:10.1038/nature06858.

20. Ekman P, Friesen WV. *Pictures of Facial Affect*. Palo Alto (CA): Consulting Psychologists Press; 1975. doi:citeulike-article-id:4270156.

21. Nikolova YS, Koenen KC, Galea S, et al. Beyond genotype: serotonin transporter epigenetic modification predicts human brain function. *Nat Neurosci*. 2014;17(9):1153-1155. doi:10.1038/nn.3778.

22. Hoaglin DC, Iglewicz B. Fine-tuning some resistant rules for outlier labeling. *J Am Stat Assoc*. 1987;82(400):1147-1149. doi:10.1080/01621459.1987.10478551.

23. Hubert M, Vandervieren E. An adjusted boxplot for skewed distributions. *Comput Stat Data Anal*. 2008;52(12):5186-5201. doi:10.1016/j.csda.2007.11.008.

24. Seo S. A review and comparison of methods for detecting outliers in univariate data sets. *Grad Sch Public Heal*. 2006. http://d-scholarship.pitt.edu/7948/.

25. Preacher KJ, Rucker DD, Hayes AF. Addressing moderated mediation hypotheses: Theory, methods, and prescriptions. *Multivariate Behav Res*. 2007;42(1):185-227. doi:10.1080/00273170701341316.