



Reliability of neural food cue-reactivity in participants with obesity undergoing bariatric surgery: a 26-week longitudinal fMRI study

Patrick Bach^{1,2} · Martin Grosshans¹ · Anne Koopmann^{1,2} · Peter Kienle³ · Georgi Vassilev⁴ · Mirko Otto⁴ · J. Malte Bumb^{1,2} · Falk Kiefer^{1,2}

Received: 12 August 2020 / Accepted: 24 November 2020 / Published online: 17 December 2020
© The Author(s) 2020

Abstract

Obesity is highly prevalent worldwide and results in a high disease burden. The efforts to monitor and predict treatment outcome in participants with obesity using functional magnetic resonance imaging (fMRI) depends on the reliability of the investigated task-fMRI brain activation. To date, no study has investigated whole-brain reliability of neural food cue-reactivity. To close this gap, we analyzed the longitudinal reliability of an established food cue-reactivity task. Longitudinal reliability of neural food-cue-induced brain activation and subjective food craving ratings over three fMRI sessions (T0: 2 weeks before surgery, T1: 8 weeks and T2: 24 weeks after surgery) were investigated in $N = 11$ participants with obesity. We computed an array of established reliability estimates, including the intraclass correlation (ICC), the Dice and Jaccard coefficients and similarity of brain activation maps. The data indicated good reliability ($ICC > 0.6$) of subjective food craving ratings over 26 weeks and excellent reliability ($ICC > 0.75$) of brain activation signals for the contrast of interest (food > neutral) in the caudate, putamen, thalamus, middle cingulum, inferior, middle and superior occipital gyri, and middle and superior temporal gyri and cuneus. Using similarity estimates, it was possible to re-identify individuals based on their neural activation maps (73%) with a fading degree of accuracy, when comparing fMRI sessions further apart. The results show excellent reliability of task-fMRI neural brain activation in several brain regions. Current data suggest that fMRI-based measures might indeed be suitable to monitor and predict treatment outcome in participants with obesity undergoing bariatric surgery.

Keywords Food cue-reactivity · Reliability · Intraclass correlation · fMRI · Dice · Jaccard

J. Malte Bumb and Falk Kiefer contributed equally to this work.

Supplementary Information The online version contains supplementary material available at <https://doi.org/10.1007/s00406-020-01218-8>.

✉ Patrick Bach
patrick.bach@zi-mannheim.de

¹ Department of Addictive Behavior and Addiction Medicine, Medical Faculty Mannheim, Central Institute of Mental Health, Heidelberg University, J5/68159 Mannheim, Germany

² Feuerlein Center on Translational Addiction Medicine (FCTS), University of Heidelberg, Heidelberg, Germany

³ Department of Surgery, Theresienkrankenhaus, Mannheim, Germany

⁴ Department of Surgery, University Medical Center Mannheim, Heidelberg University, Mannheim, Germany

Introduction

Obesity affects more than 650 million people worldwide [43]. Overweight has been identified as a major cause of cardiovascular diseases, diabetes, musculoskeletal disorders as well as several types of cancer [5]. The assessment of behavioral and neural responses towards food cues has received some interest in the last decade as a tool to investigate the neurobiological basis of obesity [18, 26]. A recent meta-analysis on food cue-reactivity concluded that across 45 published reports the overall effect of food cue-reactivity and craving on outcomes in patients was of medium size ($r = 0.3$) with a large variability across studies. Authors concluded that food cue exposure and the experience of craving have a significant influence on and contribute to eating behavior and weight gain [6]. Functional magnetic resonance imaging (fMRI) was used to identify the neural correlates of food craving, food perception, and food intake. Structures implicated in food-intake regulation include the anterior

insula, inferior frontal and orbitofrontal cortices, the medial temporal cortex with the amygdala and parahippocampus, as well as the nucleus accumbens, and visual cortices [2]. In the last years, efforts were undertaken to establish neural predictors for eating behavior and treatment response. Some studies reported significant associations between neural responses to food cues and weight loss during treatment, but overall, the studies report heterogeneous findings [22]. The inconsistencies in study results demand for an investigation of the reliability and robustness of the applied food cue-reactivity task, because the possibility to establishing meaningful and robust associations between neural brain responses during food cue presentation and any behavioral or clinical variable critically depends on the reliability of the investigated task-fMRI brain activation. Previous studies have demonstrated substantial variability in findings of food cue-reactivity studies. Although brain responses to visual food cues in participants with obesity have been found to have relatively good mean-level reproducibility, they had poor within-subject test–retest reliability. Several factors were associated with the heterogeneity in findings, including different expression of the fat mass and obesity-associated genes (e.g. FTO) [25, 36, 44], fasted state vs. glucose ingestion prior to fMRI [20] and divergent characteristics of the individual study designs, including the structure, timing and stimuli used during the food cue-reactivity fMRI task. Furthermore, there are clear individual differences in food preferences that were associated with additional variance across studies [42]. Additionally, small sample sizes and a lack of power were related to inconsistencies between studies [8]. Moreover, a study comparing the results of 70 different teams analyzing the same dataset, revealed significant variability in the analysis of the same fMRI food cue-reactivity dataset depending on the researchers decision to use a certain the statistical software (e.g. SPM vs. FSL vs. AFNI) or statistical method (parametric vs. non-parametric) as well as the applied smoothing kernel [7]. The results highlight the need for better standardization of the food stimuli and fMRI task designs and the additional data that are collected on participant's state (hunger, mood, hormones etc.) and personal characteristics that may be used to control for confounding effects in the analyses. The aforementioned findings emphasize the importance of establishing standardized food cue-reactivity paradigms, study protocols and analysis workflows. To this end, guidelines for good practice in food cue-reactivity neuroimaging studies were proposed. According to these guidelines, researchers planning fMRI studies should take special care to: power calculation, hunger state and related factors, personal characteristics, the selection of food-related stimuli, setting well-considered statistical thresholds for whole-brain analyses, minimizing the risk of movement artifacts, analysis of prospective designs as well as predictive modelling. Moreover, the authors suggest to

pre-register planned studies and to share the data obtained [40]. In doing so, it would be possible to ensure reproducibility of results across cue-reactivity studies [40].

To date, there is no study that investigated whole-brain reliability of food cue-induced brain activation. To our knowledge, only a single fMRI study investigated the longitudinal reliability of extracted mean brain activation during food cue processing over a mean period of 18 days (3–35 days), which is short considering follow-up periods of clinical studies that run over months. Additionally, reliability was only assessed in a selected range of a priori defined regions of interest (bilateral insula, amygdala, orbitofrontal cortex, caudate and putamen) [9]. The authors reported that in their dataset, only the left orbitofrontal cortex response showed fair reliability, while all other regions of interest showed poor reliability. The authors also stated that the large inter-individual range of days between the two assessment sessions might have limited reliability in their study. Additionally, previous research highlighted that low reliability in fMRI studies might also be associated to the computation of difference scores or difference contrasts, where one condition is subtracted from the other. For example, regarding the food cue-reactivity tasks, it is common to subtract the brain activation during food picture blocks from activation during neutral picture blocks. However, in the case of a high correlation between the constituting conditions of a difference score, the resulting reliability of that score is limited, because much of the shared “true” variance is removed, while the measurement errors are added [23, 35]. To date, however, no study investigated whole-brain reliability of food-cue-induced brain responses over a longer period of time and, importantly, no study to date investigated reliability in samples of patients undergoing surgery. This, however, seems relevant to the ongoing efforts to establish predictors and biomarkers for treatment efficacy in obesity. In this context, it is necessary to determine the reliability of food cue-reactivity in clinical populations undergoing treatment, because only this way the robustness and suitability of cue-reactivity as a biomarker in obesity can be assessed. Hence, we conducted our analyses in a clinical population undergoing surgery, as this sample reflects a sample for whom biomarkers should be established to predict and monitor treatment outcomes using fMRI biomarkers. Hence, we set out to assess the reliability of neural food cue-reactivity in a longitudinal dataset of individuals with obesity over three neuroimaging assessments that were scheduled 2 weeks before bariatric surgery, and 8 and 24 weeks after surgical intervention. We used an unrestricted whole-brain approach and a set of complementary measures for fMRI reliability, aimed at determining the global and local reliability of the difference contrast (food-neutral) and of the constituting food and neutral picture conditions. Additionally, we compared the reliability of food cue-reactivity to the reliability

of commonly applied subjective craving measures that were measured during the fMRI session.

Subjects and methods

Participants

Current analyses were conducted on a dataset of $N=11$ individuals with obesity of whom fMRI task data was available for three time points and that were part of a larger longitudinal clinical study, including a total of $N=26$ participants with obesity, of whom, however, only the $N=11$ participants met the inclusions criteria for undergoing fMRI scanning (e.g. absence of metal implants, claustrophobia and waist circumference < 160 cm (due to the scanner diameter). The clinical data of the of the whole study group are reported elsewhere. In short, patients showed a percent total weight loss after surgery (%TWL) from T0 to T2 of 23.8%TWL after Roux-en-Y gastric bypass ($n=21$) and 12.7%TWL after sleeve gastrectomy ($n=5$) with no significant difference between both procedures ($p=0.126$). There were also significant reductions of resting heart rate, fasting plasma glucose levels and depressive symptoms (all $p < 0.001$). Only individuals with obesity that already decided to receive bariatric surgery were recruited for this study. The study procedure was approved by the local ethics committee and all participants provided written informed consent.

Individuals with obesity undergoing fMRI had to meet the following inclusion criteria: (i) age between 18 and 65 years, (ii) BMI (kg/m^2) > 35 (i.e. \geq grade 2 obesity), (iii) a waist circumference < 160 cm (limited by scanner diameter), (iv) the capacity to give informed consent, (v) no history or current diagnosis of any psychiatric, neurological, neoplastic or untreated endocrine illnesses (with the exception of nicotine addiction), and no current intake of any centrally acting psychoactive or anti-obesity medications (i.e. sedatives, antipsychotics, including long-acting injectable antipsychotics, antidepressants, opioid analgesics as well as DPP (dipeptidyl peptidase IV) inhibitors and GLP (Glucagon-like peptide)-1 antagonists, (vi) all participants with a history of surgical interventions in the gastrointestinal system or contraindications to fMRI scanning (e.g. metal implants), and pregnant or breast-feeding females were excluded.

Twenty-six individuals (17 females and 9 males, mean age 41 ± 12 years, mean BMI $46 \pm 6 \text{ kg}/\text{m}^2$) were eligible for analyses (demographics, bariatric surgery, blood analyses as well as behavioral data) and included in the study. Of these 26 individuals, 21 received Roux-en-Y gastric bypass and 5 sleeve gastrectomy. Imaging data could be obtained for 11 obese individuals (10 individuals received Roux-en-Y gastric bypass and 1 sleeve gastrectomy; 15 individuals had to be excluded due to the fact that they did not fit the scanner.

Procedures

T0 (Two weeks before bariatric surgery)

During the first assessment session, sociodemographic data, information on internal and neurological disorders, as well as information on eating habits was collected. In addition, participants were screened for any psychiatric comorbidities using the Structured Clinical Interviews for DSM-IV, SKID-I, [45]. Additionally, a urine drug screening, and in females a pregnancy test was conducted.

fMRI scanning was performed between noon and 3 PM. All participants received a standardized breakfast of 500 kcal (2093 kJ) 6 h before fMRI scanning and did not eat until the scanning. Subsequently, participants completed a series of questionnaires including the Beck Depression Inventory (BDI, [1], the Fagerstrom Test for Cigarette Dependence (FTCD [12] as well as the Yale Food Addiction Scale (YFAS) [16].

T1 and T2 (Eight and 24 weeks after bariatric surgery)

At both time points, participants were examined medically, urine drug screenings, and in females a pregnancy test were performed. Moreover, possible changes in medication were documented. MRI measurements were performed at both time points using the same procedures and tasks as during the first scanning session.

Imaging procedure

fMRI food cue-reactivity task

All patients included in the current analyses underwent three different imaging sessions. During these sessions, patients laid in the scanner wearing MRI-compatible goggles, on which sets of visual food and neutral stimuli were presented using a block design. The task consisted of a total of 18 blocks of food stimuli and 12 blocks of neutral stimuli. Each block comprised of a series of five food or neutral pictures. Food stimuli were further divided in three categories: salty high-calorie, sweet high-calorie, low-calorie, yielding six blocks for each category. All stimuli were shown for 4 s (i.e. 20 s per block) in a pseudo-randomized order. Participants were instructed to closely watch each picture and were informed that they will be asked to rate their subjective craving. In-between each picture block, patients were asked to rate their current craving for food on a visual analogue scale (VAS) that ranged from 0—“very weak” to 100—“very strong”. The fMRI took 18 min. Food stimuli chosen were rated according to their ability to induce food craving by 44

voluntary participants at our institution [17] and neutral cues were taken from the International Affective Picture Series [28].

fMRI acquisition and pre-processing

A total of 453 images T2*-weighted, echo planar images covering the entire brain were acquired during the food cue task using a 3-T whole-body tomography scanner (MAGNETOM Trio with TIM technology; Siemens). Imaging parameters were: repetition time = 2.41 s, echo time = 25 ms, flip angle = 80°, number of slices = 42, slice thickness = 2 mm, voxel-gap = 1 mm, voxel dimensions = 3 × 3 × 3 mm³, field of view = 192 × 192 mm², in-plane resolution = 64 × 64. The short echo time and the 30° flip angle to anterior commissure–posterior commissure orientation was chosen to minimize susceptibility artefacts. Stimuli were presented using Presentation software (version 9.9, Neurobehavioral Systems Inc.) and MRI-compatible goggles (MRI Audio/Video Systems; Resonance Technology Inc., CA).

Functional-imaging data were processed and analyzed using SPM8 and SPM12. The first five scans were excluded from imaging analyses to avoid any artefacts caused by the effects of magnetic saturation. All images were realigned spatially (movement was considered excessive with > 2 mm translation or > 2° rotation), normalized to a standardized EPI template from MNI (Montreal Neurological Institute, Quebec, Canada), and smoothed using an isotropic Gaussian kernel for group analyses (full width at half maximum: 8 mm).

Food cue-reactivity imaging data were analyzed by modelling the different task conditions (food with the subcategories salty high-calorie, sweet high-calorie, low-calorie and neutral) as explanatory variables within a general linear model in SPM implementing the movement parameters as nuisance variables. Individual contrast images (food cues > neutral cues) were computed for each individual and then included into following second-level analyses in SPM. Nicotine consumption (categorical) was considered as covariate, because previous work indicated that nicotine modulates food cue-reactivity [27]. To satisfy a family-wise error rate correction of pFWE < 0.05, we determined a combined height ($p < 0.001$) and extent ($k \geq 103$) threshold by running 10,000 Monte Carlo simulations using AlphaSim as implemented in the Neuroelf analysis package (www.neuroelf.net) [4], (estimated smoothness was $x/y/z = 10.13/9.86/10.33$ mm) [11].

Reliability analyses

We investigated the reliability of subjective food craving ratings (i.e. mean craving for food–mean craving for neutral

stimuli during the 1st, 2nd and 3rd assessment session to correspond to the fMRI task contrast “food–neutral”) over the three fMRI sessions by computing the intraclass correlation coefficients using a two-way, mixed effects model in IBM SPSS (version 25.0). Additionally, whole-brain longitudinal reliability of individual brain responses to food stimuli over the three imaging sessions by computing measures of local and global reliability using the fmrelt toolbox for SPM12 by Kroemer, Frohner and colleagues [15] (<https://github.com/nkroemer/reliability>). Analyses were conducted on the whole brain without a-priori restrictions to specific regions of interest.

Jaccard and Dice coefficients

We computed the modified Jaccard coefficient, a common measure in fMRI reliability studies between the three different time points for the difference contrast food > neutral and the constituting contrasts (i.e. food and neutral separately). It is defined as the size of the intersection divided by the size of the union of the voxel sets and computed as follows:

$$\text{Jaccard}(A, B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$

The Jaccard coefficient can be interpreted as the percentage of overlapping significant voxels above a predefined statistical threshold (e.g. $p < 0.001$) within all significant voxels [24, 31].

Additionally, we computed the Dice coefficient for the three different contrasts and scanning time points. It is calculated as the number of super-threshold voxels that overlap between sessions divided by the average number of significant voxels across sessions:

$$\text{Dice}(A, B) = \frac{2|A \cap B|}{|A| + |B|}$$

The Dice coefficient was introduced to assess the overlap of significant fMRI clusters between scans. It has become an established measure of fMRI data reliability [37]. Both coefficients have values from 0 (“no overlap”) to 1 (“perfect overlap”) between significant super-threshold voxels. Both measures are, however, limited by the missing consensus on specific values or cut-offs that would differentiate between “poor” and “good” values [3]. Additionally, the magnitude of both coefficients depends on the statistical threshold used to define what is “active”. Studies showed that the reliability of the cluster overlap method decreases, when the significance threshold is increased [10, 38]. In the current analyses, we, therefore, applied a commonly used threshold of $p < 0.001$. Resulting values were imported into the IBM SPSS Statistics software (version 25.0) for further analyses using a repeated measure analysis of variance (ANOVA)

model with the factors time (1st, 2nd, 3rd) assessment and task contrast (food, neutral, food > neutral).

Similarity

Second, we calculated the within- and between-subject similarity of the fMRI activation maps using the fmrelt toolbox [15]. Similarity in this context is defined as the resemblance of two activation patterns based on the alignment of high versus low brain activation values across the brain between- and within-subjects (for details see Frohner et al. 2019). The resulting coefficients are correlation coefficients that range from ‘perfect’ inverse relationship (– 1.00) to a ‘perfect’ direct relationship (1.00). It was suggested that individuals can be successfully identified by their neural activation patterns, if the within-subject similarity exceeds all between-subject association coefficients of the same participant [13, 15]. An advantage of this procedure is that it does not require an a-priori (and potentially arbitrary) statistical threshold.

Intraclass correlation (ICC)

Third, we estimated voxel-wise reliability of brain activation patterns by computing the intraclass correlation (ICC) coefficients between all three fMRI sessions. The ICC is used to assess whether the magnitude of activation in each voxel of the brain is stable from test scan to retest scan. Previous work suggested that this measure might be more stringent than other fMRI reliability measures, as it also requires near zero values to be stable over time [3]. It was suggested that the ICC(3,1) variant is most appropriate for assessing longitudinal fMRI datasets [34]. Mathematically, this coefficient sets within-subject variance (σ^2_{within}) in relation to between-subject variance ($\sigma^2_{\text{between}}$). We used the ICC(3,1)-type to assess voxel-wise reliability [39], defined as:

$$\text{ICC} = \frac{(\sigma^2_{\text{between}} - \sigma^2_{\text{within}})}{(\sigma^2_{\text{between}} + \sigma^2_{\text{within}})}.$$

According to Fleiss (1986), ICC coefficients lower than 0.4 represent poor reliability, ICCs between 0.4 and 0.75 represent fair (<0.6)-to-good (>0.6) reliability, and ICCs higher than 0.75 represent good-to-excellent reliability [14]. We calculated ICC coefficients for every brain voxel to allow identification of brain regions that show high reliability without restriction to predefined regions of interest. However, we were aware that much of the (un-thresholded) brain activation might be unrelated to food cue task and hence would not replicate in its magnitude, resulting in a low overall ICC value. Therefore, we generated thresholded ICC brain maps, to identify brain areas that show good-to-excellent (ICC > 0.75) reliability and we computed

additional atlas-based mean ICC values for a standard set of anatomical brain regions (see below).

Spearman’s correlation

To assess whether reliability of the common difference contrast food > neutral might be limited by a high correlation between the constituting conditions, we computed the voxel-wise Spearman’s correlation coefficients between the three food image categories (i.e. sweet, high caloric, low caloric) and the neutral condition using the fmrelt toolbox.

Computation of atlas-based summary measures

In accordance to previous work [15], we computed the mean ICC for $N=120$ anatomical regions specified in the Automatic Anatomic Labeling (AAL) atlas [41]. The additional atlas-based summary intended to facilitate the assessment of local differences in reliability and identify reliable anatomical ROIs for future analyses. ICC values were extracted using the ROI data extraction routine of the MarsBar software package (<http://marsbar.sourceforge.net/>) and was imported into SPSS (IBM SPSS Statistics version 25.0) for further analyses.

Group-level fMRI task activation

On a group level, imaging data for every single time point (e.g. 1st, 2nd, 3rd assessment) were analyzed using a one-sample t test to assess the robustness of task main effects (i.e. between condition effects) on group-level brain activation and to determine brain areas that show higher brain activation in response to food cues, compared to neutral cues (contrast: food–neutral). Additionally, we performed analyses of changes in food cue-induced brain responses over time, by setting up a flexible factorial model with the within subject factor time (i.e. 1st, 2nd, 3rd assessment) and the covariates BMI at baseline, surgery type and smoking status. In order to satisfy a family-wise error rate correction of $p\text{FWE} < 0.05$, we determined a combined voxel-wise [$p < 0.001$] and cluster-extent-threshold [$k \geq 103$] by running 10,000 permutations by Monte Carlo simulations (the estimated smoothness was $x/y/z = 10.13/9.86/10.33$ mm) using the Neuroelf analysis package (www.neuroelf.net) (Bennett et al. 2009) [4].

Results

Sample characteristics

Demographical, clinical and psychometric data are depicted in Table 1.

Table 1 Demographic and clinical characteristics of obese study participants that underwent three imaging assessments at T0=2 weeks prior to surgery, T1=8 weeks after surgery and T2=24 weeks after surgery ($N=11$)

$N=11$ participants with obesity	Absolute numbers	Relative proportions (%)
Sex (male/female)	3/8	27.3/72.7
Smoking status (non-smoking/ < 10 cig. per day/ > = 10 cig. per day)	7/2/2	63.6/18.2/18.2
	Mean	SD
Age (years)	41.18	10.1
Height (m)	1.68	0.1
Weight (kg)	128.78	17.1
BMI	45.40	4.7
BDI (total score)	9.45	4.6

Group-level food cue-induced brain activation

Group-level analyses of brain activation demonstrated significant food cue-induced brain activation (contrast: food > neutral) in parts of the frontal and orbitofrontal cortex, the occipital and parietal gyri, the cuneus, calcarine, the lingual gyrus, as well as the caudate, putamen, thalamus and insula (see Table 2). On the other hand, no significant brain activation was detected during presentation of neutral pictures compared to food pictures (contrast: neutral > food). Whole analyses of longitudinal changes in brain responses towards food cues over assessment sessions before and after surgery showed no main effect of time on brain response towards food cues and post-hoc comparisons between separate assessment time points did not surpass the predefined statistical threshold.

Reliability analyses

Food craving ratings

Analyses indicated good reliability of the mean subjective food craving ratings during fMRI across the different

assessment sessions (ICC [1, 3] = 0.611, $p=0.002$). Food cues induced higher craving values compared to neutral cues throughout all three assessment sessions. There was a significant reduction in the magnitude of reported food craving over the trial period from baseline ($M=45.195$, $SD=23.443$) to T1 ($M=18.550$, $SD=39.917$) that remained stable until T2 ($M=32.450$, $SD=25.972$, $F_{(2,18)}=4.301$, $p=0.032$).

Jaccard coefficient

Mean Jaccard coefficients for the comparisons of the different time points are displayed in Table 3. Repeated measures ANOVA showed a significant main effect of contrast image category (neutral, food and food > neutral) ($F_{(2,20)}=83.806$, $p<0.001$) on the magnitude of the Jaccard indices. Post hoc analyses demonstrated lower Jaccard coefficients for the difference contrast condition (food > neutral) compared to both constituting conditions (food and neutral, $p<0.001$). There was no main effect of time on the magnitude of the Jaccard coefficients (i.e. whether we compared to 1st to 2nd or 3rd scanning session, $F_{(2,20)}=0.152$, $p=0.860$).

Table 2 Brain depicting higher brain response to visual food cues compared to neutral cues (contrast: food > neutral, combined voxel-wise [$p<.001$] and cluster-extent-threshold [$k>103$ voxel]), corresponding to pFWE < .05)

Side	Lobe	Brain areas	Cluster size (voxel)	MNI coordinates (x, y, z)	t_{\max}
R and L	Occipital	Superior, Middle and Inferior Occipital Gyrus, Calcarine, Cuneus, Fusiform Gyrus, Lingual Gyrus	7081	32 - 76 - 14	21.9
R	Parietal	Inferior Parietal Gyrus, Angular Gyrus	133	32 - 68 54	9.6
L	Occipital, Parietal	Superior and Middle Parietal and Occipital Gyrus	275	- 24 - 60 44	8.9
L		Putamen, Insula	129	- 40 - 6 10	8.7
L	Parietal	Inferior Parietal Gyrus, Postcentral Gyrus, Supramarginal Gyrus	142	- 48 - 24 40	8.6
R and L		Anterior and Middle Cingulate Gyrus	176	- 8 24 24	7.4
L	Frontal	Middle and Inferior Frontal Gyrus, Orbitofrontal Cortex	130	- 44 36 14	7.2
R		Caudate, Thalamus	104	14 - 4 12	6.9

Table 3 (A) Dice and (B) Jaccard coefficients for the three task contrasts (food > neutral, food and neutral), illustrating the proportion of overlapping significant voxels between the different fMRI sessionsat T0=two weeks prior to surgery, T1=eight weeks after surgery and T2=twenty-four weeks after surgery (whole-brain threshold of $p < 0.001$ for defining super-threshold activation)

Comparison of sessions	Session 1 and 2			Session 1 and 3			Session 2 and 3		
(a) Dice coefficients									
Contrast	Food > Neutral	Food	Neutral	Food > Neutral	Food	Neutral	Food > Neutral	Food	Neutral
Mean	0.2743***	0.6763	0.7103	0.2049***	0.7181	0.7260	0.2218***	0.6921	0.6790
SD	0.2036	0.2160	0.2067	0.1599	0.0763	0.0762	0.1918	0.2187	0.2106
(b) Jaccard coefficients									
Contrast	Food > Neutral	Food	Neutral	Food > Neutral	Food	Neutral	Food > Neutral	Food	Neutral
Mean	0.1744***	0.5400	0.5772	0.1222***	0.5651	0.5750	0.1375***	0.5596	0.5406
SD	0.1443	0.1970	0.1830	0.0996	0.0911	0.0935	0.1316	0.2024	0.1853

SD standard deviation

***Significant difference at $p < 0.001$ between the contrast condition food > neutral and each of the other two conditions (food and neutral)

Dice coefficient

The mean Dice coefficients for the comparisons of the different fMRI sessions are depicted in Table 3. Analyses demonstrated a significant main effect of contrast image category (neutral, food and food > neutral) ($F_{(2,20)} = 77.102$ $p < 0.001$) on the magnitude of the Jaccard indices. Post hoc analyses demonstrated lower Jaccard coefficients for the difference contrast condition (food > neutral) compared to both constituting conditions (food and neutral, $p < 0.001$). There was no main effect of time on the magnitude of the Jaccard coefficients (i.e. whether we compared to 1st to 2nd or 3rd scanning session, $F_{(2,20)} = 0.208$ $p = 0.814$).

ICC

Comparisons of ICC coefficients between the different fMRI sessions (1st, 2nd, 3rd) indicated that several regions showed good to excellent reliability (i.e. ICC > 0.75) across all sessions (see Fig. 1). These regions included the bilateral caudate and left putamen, parts of the right thalamus and middle cingulum, as well as parts of the bilateral inferior, middle and superior occipital gyri (brodmann areas BA 7/17/18/19/39) and parts of the bilateral middle and superior temporal gyri (BA 20/21/22/37) and in addition parts of the bilateral cunei, lingual gyri and calcarine (see Fig. 1). These patterns appeared to be relatively stable across all session time points, supporting the stability of the observed findings.

In a second step, we determined the mean ICC for a standard set of $n = 120$ anatomical regions of interest defined in the aal atlas. As expected based on the patterns of voxel-wise ICC values (i.e. good to excellent reliability only in parts of the anatomical region), the mean overall ICC for the separate regions did not exceed the voxel-wise values. However, several anatomical regions of interest masks showed good or fair reliability (see supplementary Table S1), specifically

the bilateral inferior, middle and superior occipital gyri ROIs showed good overall reliability (> 0.6) and the several other regions showed fair reliability (> 0.4) Left putamen, bilateral caudate, left amygdala, bilateral lingual gyri, right fusiform gyrus, bilateral calcarine, bilateral cunei, posterior cingulate, right middle temporal gyrus, bilateral middle frontal gyri, right superior medial gyrus, left superior parietal gyrus and angular gyrus. The ICC maps underlying the presented results are provided on Neurovault.org (<https://identifier.s.org/neurovault.collection:9026>).

Spearman's rho

We computed spearman's rho coefficients between the food category contrast maps and the neutral contrast maps to assess whether there is a high correlation between the constituting conditions (food and neutral), which would reduce the maximum possible reliability of the difference contrast (food-neutral), due to elimination of shared variance during performing the subtraction. Results demonstrate a substantial correlation between the all three food stimuli category contrast maps and the neutral stimuli contrast maps ($\rho_{\text{sweet-neutral}} = 0.49$, $SD = 0.29$, $R^2 = 0.24$, $\rho_{\text{low-neutral}} = 0.42$, $SD = 0.33$, $R^2 = 0.17$, $\rho_{\text{high-neutral}} = 0.42$, $SD = 0.33$, $R^2 = 0.17$). This indicates that both food and neutral conditions share about 17–24% of their variance. A part of this variance is removed by subtracting both conditions, which results in lower reliability of the difference contrast [23].

Similarity

The analyses of similarity between activation maps for the difference contrast (food > neutral) showed a gradual decrease of within-subject similarity over comparisons between fMRI sessions with increasing time between the

Brain regions with ICC values > 0.75 (“good”) for contrast food > neutral and N = 11 obese patients

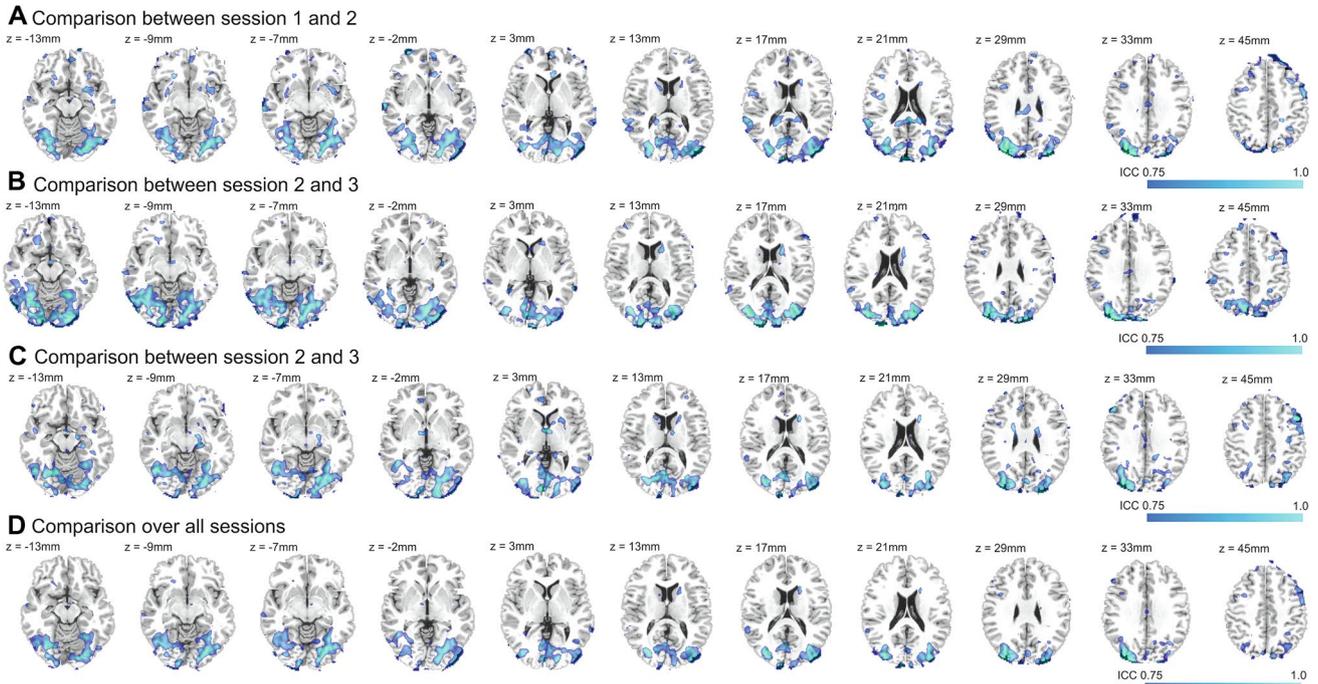


Fig. 1 Depiction of brain areas that show good to excellent reliability for the difference contrast food-neutral (Intraclass correlation [ICC]>0.75) for the comparisons between: **a** session 1 and 2 (i.e.

2 weeks prior to surgery and 8 weeks after surgery), **b** session two and three (i.e. 2 weeks prior to surgery and 24 weeks after surgery), **c** session 1 and three and **d** over all sessions

respective sessions (i.e. higher within-subject similarity between T0 and T2 that were 10 weeks apart vs. T2 and T3 that were 16 weeks apart). This reflected in lower t values for the comparisons between within-subject and between-subject similarity for the respective sessions and lower mean similarity values ($r_{T0_T1}=0.37$, $t_{T0_T1}=5.14$, $p<0.001$, $r_{T1_T2}=0.32$, $t_{T1_T2}=3.82$, $p<0.05$, $r_{T1_T3}=0.29$, $t_{T1_T3}=3.01$, $p<0.05$). The difference between within and between-subject similarity is visible in the matrices and cumulative distribution functions for within- and between-subject similarity in Fig. 2. The proportion of patients that could be re-identified based on their neural brain activation (i.e. the magnitude of within-subject similarity exceeded all between-subject similarity values). While about 73% of the patients could be re-identified between T0 and T1, this number dropped when comparing longer time periods between T1–T2 (64%) and T0–T2 (45%, see Fig. 2).

Discussion

The purpose of this study was to investigate the longitudinal reliability of the different task contrasts an established food cue-reactivity task. ICC values indicated good to excellent reliability of brain activation, captures by the common

difference contrast food vs. neutral, in a range of brain areas (i.e. the mesolimbic system with putamen and caudate, as well as parts of the frontal and occipital cortices) over a time period of 26 weeks. In addition the reliability of food cue-induced brain activation in these brain regions, indexed by the difference contrast food vs. neutral, outperformed the reliability of subjective food craving (i.e. craving during food blocks vs. neutral blocks) that was measured concurrently during fMRI using visual analogue scales. Still, it should be noted that local reliability did not surpass the threshold for good reliability in all areas of the mesocorticolimbic system, which were implicated in processing food cues [33]. Furthermore, Jaccard and Dice coefficients, which provide estimates for the replicability of significant activation clusters across the whole brain, indicated that only a small proportion of activation could be replicated, when investigating the difference contrast (food > neutral). This stood in sharp contrast to the results for the constituting task contrast conditions food vs. baseline and neutral vs. baseline separately. For these two contrast conditions, Jaccard and Dice coefficients showed that more than 50% of the super-threshold clusters could be replicated during the other assessment sessions. This indicates that the global reliability of the common difference contrast food vs. neutral is limited. Several reasons might account for these findings. In previous studies,

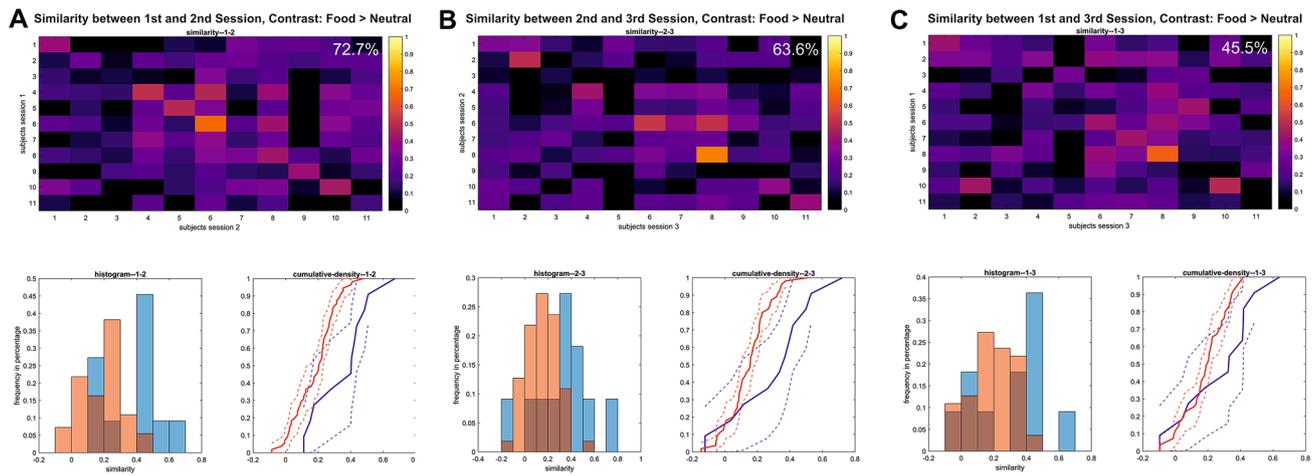


Fig. 2 Similarity maps (upper row) and empirical cumulative distribution functions (lower row—red lines: between-subject similarity, blue lines: within-subject similarity) for the contrast food–neutral and comparisons between **a** 1st and 2nd fMRI session, **b** 2nd and 3rd fMRI session and **c** 1st and 3rd fMRI session. The diagonal of each color matrix represents the within-subject similarity values. Re-identification of a subject based on the neural activation map is affirmed the within-subject similarity value (diagonal) exceeds all between-subject association coefficients of the same participant (i.e.

similarity values in the respective row of the matrix). Higher within-subject similarity is also illustrated by a right-shift of the cumulative density functions for the within-subject similarity values (blue lines) relative to the between-subject similarity (red lines). Percent values in the upper right of the upper row panels represent the number of individuals that could be identified based on their brain response (i.e. within-subject similarity values exceeded all between-subject similarity values for the respective participant [rows in matrix])

Infantolino and colleagues (2018) argued that the correlation between the constituting contrast conditions of a difference contrast place a limit on the reliability of the resulting difference measure, because in this case, large proportions of the shared and potentially true variance are eliminated by subtracting both constituting task conditions. The authors sustained their argument with data on the difference contrast between face- and shape-matching trials of a so-called faces paradigm, where the constitution shape and face conditions correlated to 0.97 [19, 23]. Other fMRI studies that also computed difference contrasts as the measure of interest, reported higher reliability of brain activation that was mirrored by an only modest correlation between the constituting conditions [30]. Current data show a moderate correlation between the food and neutral contrast images with a shared variance of about 24%. This supports the notion that the global reliability of the difference contrast (food vs. neutral) in the current dataset is limited by the correlation between the constituting conditions, which results in an elimination of proportions of the shared variance. The similarity analyses indicated that the capacity to identify individual individuals based on their individual brain activation pattern during the food vs. neutral contrast fades, when time periods between sessions increase. This was an expected finding and suggests that in the case of food cue-reactivity, follow-up fMRI scans should not be scheduled too far apart, when one intends to yield high reliability.

The only other previous study specifically investigating reliability of food cue-reactivity used a pre-selected range of ROIs (insula, putamen, amygdala, orbitofrontal cortex, caudate) and reported overall poor reliability in these ROIs. Several reasons might have accounted for the differences between this and the current study. The study by Drew Sayer et al. [9] used a different fMRI task design. The number of blocks of neutral and food stimuli per run was markedly lower (i.e. 3 and 3) compared to the task that was used as a basis for current analyses. Fewer data points per subject might, however, lead to less robust estimates of the individuals “true” mean value, e.g. brain response. Additionally, the study did not investigate voxel-wise reliability, but instead extracted brain activation estimates from predefined regions of interest and focused on the ICC as only an estimate for reliability. The use of the local maxima that were detected in the group level analyses as center of these ROIs, might have biased results. Studies have shown that a robust effect on the group level does not indicate stability or reliability of within-subject effects and might also be influenced by outliers [23]. Hence, the focus on these specific ROIs that only covered a diameter of 3 mm around the activation maximum, might have limited the possibility to identify regions with robust reliability. Current atlas-based summaries support the notion that the areas under investigation, specifically the caudate, putamen and amygdala show at least moderate reliability, when using the fMRI task of the current study.

Multiple studies intended to determine predictors for successful weight loss after bariatric surgery and establish neural “biomarkers” [21, 32]. As reliability is a prerequisite for any measure that could potentially serve as “biomarker”, current results could inform future studies and support the notion that neural responses to food cues in a selected range of brain areas might indeed meet the requirements for a potential predictor of treatment outcome.

Strengths and limitations

We investigated a specific block-design food cue-reactivity task that was used and validated in previous work by our group [17]. Hence, our results may be generalized for food cue-reactivity tasks that incorporate different picture sets or a different task design. Still, the convergence of the different reliability estimates supports the robustness of the findings and the applied methods. We also acknowledge that other methods for the estimation of fMRI reliability exist (e.g. support vector machine learning) and might be informative. We investigated the reliability in a clinical population undergoing surgery. Due to the fact that reliability depends on the population under investigation, we argue that this approach complements the investigation of healthy reference samples to assess the potential of fMRI-based markers for application in clinical populations. Still, the investigation of healthy samples and individuals with obesity are essential to yield robust estimates of reliability of food cue-reactivity without potential bias and reduction in reliability due to surgical intervention or weight loss and improve the overall precision of reliability estimates. We, however, intended to provide a conservative estimate of the reliability of food cue-reactivity, because we acknowledge that statistical control might not be feasible and is also arbitrary to a certain extent (e.g. only controlling variables that show a significant effect of time in a respective trial would lead to differences between trials). This might lead to bias in the estimating of the reliability of food cue-reactivity. It could be argued that the inclusion of patients without any treatment might be favorable with regards to yielding optimal reliability. However, we strongly advocate for testing reliability under the conditions in which the actual task is applied. When intending to use neural brain response as biomarker for monitoring e.g. treatment response, reliability of this putative biomarker should be tested under the very same conditions. It should be noted that reliability estimates, which are based on small datasets, are prone to imprecision, due to large confidence intervals and high impact of single participant data, which also accounts for the presented dataset. The complementary whole-brain analyses that compared brain responses towards food cues between the different assessment session did not yield significance, when applying a stringent whole-brain

correction for multiple testing. This result is unexpected and contrasts previous studies that showed longitudinal changes in brain response from before to after surgery [29, 46]. The lack of significant main effects of time on brain response might relate to a limited power and a stringent whole-brain threshold (e.g. previous studies applied regions of interest analyses), resulting from the small dataset. However, power analyses indicated that analyses comparing different time points yielded sufficient power (see Supplementary Figure S1). Additionally, several significant findings were derived from studies applying more liberal regions of interest analyses. Overall, the lack of substantial time effects on the extent of food cue-induced brain response in the current dataset support the notion that reliability estimates were not substantially biased by surgical intervention.

Conclusion

We could show excellent local longitudinal reliability in a range of brain areas of the reward (e.g. caudate, putamen) and food-cue-processing networks (e.g. occipital and frontal cortices) in participants with obesity from 2 weeks before, to 24 weeks after surgery. The reliability of food cue-reactivity in these areas outperformed to reliability of subjective craving measures that were measured concurrently. Our results suggest that fMRI-based measures might indeed be suitable to monitor and predict treatment outcome in participants with obesity undergoing bariatric surgery.

Acknowledgements We thank Ms. U. Schmid for proofreading and editing the manuscript.

Funding Open Access funding enabled and organized by Projekt DEAL. The current study was conducted without additional financial support from any external research bodies.

Compliance with ethical standards

Conflict of interest None.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and/or national research committee (ethics committee II Heidelberg University) and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Informed consent Informed consent was obtained from all individual participants included in the study.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article’s Creative Commons licence, unless indicated

otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>.

References

1. Beck AT, Ward CH, Mendelson M, Mock J, Erbaugh J (1961) An inventory for measuring depression. *Arch Gen Psychiatry* 4:561–571
2. Benarroch EE (2010) Neural control of feeding behavior: overview and clinical correlations. *Neurology* 74:1643–1650
3. Bennett CM, Miller MB (2010) How reliable are the results from functional magnetic resonance imaging? *Ann N Y Acad Sci* 1191:133–155
4. Bennett CM, Wolford GL, Miller MB (2009) The principled control of false positives in neuroimaging. *Soc Cogn Affect Neurosci* 4:417–422
5. Bhaskaran K, Douglas I, Forbes H, dos-Santos-Silva I, Leon DA, Smeeth L (2014) Body-mass index and risk of 22 specific cancers: a population-based cohort study of 5.24 million UK adults. *Lancet* 384:755–765
6. Boswell RG, Kober H (2016) Food cue reactivity and craving predict eating and weight gain: a meta-analytic review. *Obes Rev* 17:159–177
7. Botvinik-Nezer R, Holzmeister F, Camerer CF, Dreber A, Huber J, Johannesson M, Kirchler M, Iwanir R, Mumford JA, Adcock RA, Avesani P, Baczkowski BM, Bajracharya A, Bakst L, Ball S, Barilari M, Bault N, Beaton D, Beitner J, Benoit RG, Berkers RMWJ, Bhanji JP, Biswal BB, Bobadilla-Suarez S, Bortolini T, Bottenhorn KL, Bowring A, Braem S, Brooks HR, Brudner EG, Calderon CB, Camilleri JA, Castellon JJ, Cecchetti L, Cieslik EC, Cole ZJ, Collignon O, Cox RW, Cunningham WA, Czoschke S, Dadi K, Davis CP, Luca AD, Delgado MR, Demetriou L, Dennison JB, Di X, Dickie EW, Dobryakova E, Donnat CL, Dukart J, Duncan NW, Durnez J, Eed A, Eickhoff SB, Erhart A, Fontanesi L, Fricke GM, Fu S, Galván A, Gau R, Genon S, Glattard T, Glerean E, Goeman JJ, Golowin SAE, González-García C, Gorgolewski KJ, Grady CL, Green MA, Guassi Moreira JF, Guest O, Hakimi S, Hamilton JP, Hancock R, Handjaras G, Harry BB, Hawco C, Herholz P, Herman G, Heunis S, Hoffstaedter F, Hogeveen J, Holmes S, Hu C-P, Huettel SA, Hughes ME, Iacovella V, Iordan AD, Isager PM, Isik AI, Jahn A, Johnson MR, Johnstone T, Joseph MJE, Juliano AC, Kable JW, Kassinosopoulos M, Koba C, Kong X-Z et al (2020) Variability in the analysis of a single neuroimaging dataset by many teams. *Nature* 582:84–88
8. Button KS, Ioannidis JP, Mokrysz C, Nosek BA, Flint J, Robinson ES, Munafò MR (2013) Power failure: why small sample size undermines the reliability of neuroscience. *Nat Rev Neurosci* 14:365–376
9. Drew Sayer R, Tamer GG Jr, Chen N, Tregellas JR, Cornier MA, Kareken DA, Talavage TM, McCrory MA, Campbell WW (2016) Reproducibility assessment of brain responses to visual food stimuli in adults with overweight and obesity. *Obesity (Silver Spring)* 24:2057–2063
10. Duncan KJ, Pattamadilok C, Knierim I, Devlin JT (2009) Consistency and variability in functional localisers. *Neuroimage* 46:1018–1026
11. Eklund A, Nichols TE, Knutsson H (2016) Cluster failure: why fMRI inferences for spatial extent have inflated false-positive rates. *Proc Natl Acad Sci U S A* 113:7900–7905
12. Fagerstrom K (2012) Determinants of tobacco use and renaming the FTND to the fagerstrom test for cigarette dependence. *Nicotine Tob Res* 14:75–78
13. Finn ES, Shen X, Scheinost D, Rosenberg MD, Huang J, Chun MM, Papademetris X, Constable RT (2015) Functional connectome fingerprinting: identifying individuals using patterns of brain connectivity. *Nat Neurosci* 18:1664–1671
14. Fleiss JL (1986) The design and analysis of clinical experiments. John Wiley & sons. New York: 11–12
15. Frohner JH, Teckentrup V, Smolka MN, Kroemer NB (2019) Addressing the reliability fallacy in fMRI: similar group effects may arise from unreliable individual effects. *Neuroimage* 195:174–189
16. Gearhardt AN, Corbin WR, Brownell KD (2009) Preliminary validation of the Yale food addiction scale. *Appetite* 52:430–436
17. Grosshans M, Vollmert C, Vollstadt-Klein S, Tost H, Leber S, Bach P, Buhler M, von der Goltz C, Mutschler J, Loeber S, Hermann D, Wiedemann K, Meyer-Lindenberg A, Kiefer F (2012) Association of leptin with food cue-induced activation in human reward pathways. *Arch Gen Psychiatry* 69:529–537
18. Harding IH, Andrews ZB, Mata F, Orlandea S, Martínez-Zalacáin I, Soriano-Mas C, Stice E, Verdejo-García A (2018) Brain substrates of unhealthy versus healthy food choices: Influence of homeostatic status and body mass index. *Int J Obes* 42:448–454
19. Hariri AR, Tessitore A, Mattay VS, Fera F, Weinberger DR (2002) The amygdala response to emotional stimuli: a comparison of faces and scenes. *Neuroimage* 17:317–323
20. Heni M, Kullmann S, Veit R, Ketterer C, Frank S, Machicao F, Staiger H, Häring H-U, Preissl H, Fritsche A (2014) Variation in the obesity risk gene FTO determines the postprandial cerebral processing of food stimuli in the prefrontal cortex. *Mol Metab* 3:109–113
21. Holsen LM, Davidson P, Cerit H, Hye T, Moondra P, Haimovici F, Sogg S, Shikora S, Goldstein JM, Evins AE, Stoeckel LE (2018) Neural predictors of 12-month weight loss outcomes following bariatric surgery. *Int J Obes (Lond)* 42:785–793
22. Huerta CI, Sarkar PR, Duong TQ, Laird AR, Fox PT (2014) Neural bases of food perception: coordinate-based meta-analyses of neuroimaging studies in multiple modalities. *Obesity (Silver Spring)* 22:1439–1446
23. Infantolino ZP, Luking KR, Sauder CL, Curtin JJ, Hajcak G (2018) Robust is not necessarily reliable: from within-subjects fMRI contrasts to between-subjects comparisons. *Neuroimage* 173:146–152
24. Jaccard P (1902) Lois de distribution florale dans la zone alpine. *Bull Soc Vaudoise Sci Nat* 38:69–130
25. Karra E, O'Daly OG, Choudhury AI, Youssef A, Millership S, Neary MT, Scott WR, Chandarana K, Manning S, Hess ME (2013) A link between FTO, ghrelin, and impaired brain food-cue responsiveness. *J Clin Invest* 123:3539–3551
26. Kerem L, Hadjikhani N, Holsen L, Lawson EA, Plessow F (2019) Oxytocin reduces the functional connectivity between brain regions involved in eating behavior in men with overweight and obesity. *Int J Obes* 44:980–989
27. Kroemer NB, Guevara A, Vollstadt-Klein S, Smolka MN (2013) Nicotine alters food-cue reactivity via networks extending from the hypothalamus. *Neuropsychopharmacology* 38:2307–2314
28. Lang P, Bradley M, Cuthbert B (1999) The international affective picture system (IAPS). In: Attention CftSoEa (ed) University of Florida, Gainesville
29. Li G, Ji G, Hu Y, Liu L, Jin Q, Zhang W, Liu L, Wang Y, Zhao J, von Deneen KM, Chen A, Cui G, Han Y, Wang H, Zhao Q, Wu K, Wiers CE, Tomasi D, Leggio L, Volkow ND, Nie Y, Zhang Y, Wang G-J (2019) Reduced plasma ghrelin concentrations are associated with decreased brain reactivity to food cues

- after laparoscopic sleeve gastrectomy. *Psychoneuroendocrinology* 100:229–236
30. Luking KR, Nelson BD, Infantolino ZP, Sauder CL, Hajcak G (2017) Internal consistency of functional magnetic resonance imaging and electroencephalography measures of reward in late childhood and early adolescence. *Biol Psychiatry Cogn Neurosci Neuroimaging* 2:289–297
 31. Maitra R (2010) A re-defined and generalized percent-overlap-of-activation measure for studies of fMRI reproducibility and its use in identifying outlier activation maps. *Neuroimage* 50:124–135
 32. Ness A, Bruce J, Bruce A, Aupperle R, Lepping R, Martin L, Hancock L, Patrician T, Malley S, Selim N, Savage CR (2014) Pre-surgical cortical activation to food pictures is associated with weight loss following bariatric surgery. *Surg Obes Relat Dis* 10:1188–1195
 33. Noori HR, Spanagel R, Hansson AC (2012) Neurocircuitry for modeling drug effects. *Addict Biol* 17:827–864
 34. Ombao H, Lindquist M, Thompson W, Aston J (2016) *Handbook of neuroimaging data analysis*. CRC Press
 35. Peter JP, Churchill GA Jr, Brown TJ (1993) Caution in the use of difference scores in consumer research. *J Consum Res* 19:655–662
 36. Rapuano KM, Zieselman AL, Kelley WM, Sargent JD, Heatherton TF, Gilbert-Diamond D (2017) Genetic risk for obesity predicts nucleus accumbens size and responsiveness to real-world food cues. *Proc Natl Acad Sci* 114:160–165
 37. Rombouts S, Barkhof F, Hoogenraad F, Sprenger M, Valk J, Scheltens P (1997) Test-retest analysis with functional MR of the activated area in the human visual cortex. *Am J Neuroradiol* 18:1317–1322
 38. Rombouts SA, Barkhof F, Hoogenraad FG, Sprenger M, Scheltens P (1998) Within-subject reproducibility of visual activation patterns with functional magnetic resonance imaging using multislice echo planar imaging. *Magn Reson Imaging* 16:105–113
 39. Shrout PE, Fleiss JL (1979) Intraclass correlations: uses in assessing rater reliability. *Psychol Bull* 86:420–428
 40. Smeets PAM, Dagher A, Hare TA, Kullmann S, van der Laan LN, Poldrack RA, Preissl H, Small D, Stice E, Veldhuizen MG (2019) Good practice in food-related neuroimaging. *Am J Clin Nutr* 109:491–503
 41. Tzourio-Mazoyer N, Landeau B, Papathanassiou D, Crivello F, Etard O, Delcroix N, Mazoyer B, Joliot M (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *Neuroimage* 15:273–289
 42. Van Der Laan LN, Smeets PA (2015) You are what you eat: a neuroscience perspective on consumers' personality characteristics as determinants of eating behavior. *Curr Opin Food Sci* 3:11–18
 43. WHO (2018) Obesity and overweight. In: World Health Organization
 44. Wiemerslage L, Nilsson EK, Solstrand Dahlberg L, Ence-Eriksson F, Castillo S, Larsen AL, Bylund SB, Hogenkamp PS, Olivo G, Bandstein M (2016) An obesity-associated risk allele within the FTO gene affects human brain activity for areas important for emotion, impulse control and reward in response to food images. *Eur J Neurosci* 43:1173–1180
 45. Wittchen HU, Zaudig M, Fydrich T (1997) *Strukturiertes klinisches interview für dsm-iv (skid-i und skid-ii) - [the structured clinical interview for dsm-iv (scid-i and scid ii)]*. Hogrefe, Göttingen
 46. Zoon HF, de Bruijn SE, Smeets PA, de Graaf C, Janssen I, Schijns W, Aarts EO, Jager G, Boesveldt S (2018) Altered neural responsiveness to food cues in relation to food preferences, but not appetite-related hormone concentrations after RYGB-surgery. *Behav Brain Res* 353:194–202