eLife

Reviewed Preprint
v1 • June 23, 2026
Not revised

For correspondence:
o.goltermann@uke.de

Competing interests: No competing interests declared

Funding: See page 10

Reviewing editor: Saad Jbabdi, University of Oxford, United Kingdom

Opposing BOLD signals and oxygen metabolism largely arise from statistical uncertainty in metabolic estimates

Ole Goltermann[1,2] ✉, **Alexander Huth**[3,4], **Christian Büchel**[1,2]

[1]Institute of Systems Neuroscience, University Medical Center Hamburg-Eppendorf, Hamburg, Germany • [2]Max Planck School of Cognition, Leipzig, Germany • [3]Neuroscience Department, University of California, Berkeley, Berkeley, United States • [4]Statistics Department, University of California, Berkeley, Berkeley, United States

eLife Assessment

This manuscript provides a timely and **important** statistical re-evaluation of a paper by Epp et al., on the discordance of BOLD and CMRO2 measures. The authors present a **convincing** case based on rigorous re-analysis of the data that these previous results arise predominantly from uncertainty in measurement, rather than physiological features. These findings have implications that are of importance to all studies of brain function using BOLD FMRI.

https://doi.org/10.7554/eLife.111743.1.sa2

Abstract

Recent work by Epp et al. (2025) reported widespread voxel-wise sign discordance between task-evoked blood-oxygenation-level-dependent (BOLD) responses and estimated changes in cerebral metabolic rate of oxygen ($\Delta CMRO_2$), raising important questions about the interpretability of BOLD functional magnetic resonance imaging. Reanalysing the dataset, we found that $\Delta CMRO_2$ estimates showed substantial voxel-wise variability across participants, consistent with the noise sensitivity of model-based metabolic estimates. When this variability was taken into account, 77.2% of voxels could not be robustly classified, as $\Delta CMRO_2$ effects lacked sufficient statistical support to determine concordance or discordance. Where classification was possible, positive BOLD responses were predominantly concordant with metabolism, whereas discordance was considerably higher for negative BOLD responses. These findings suggest that the observed BOLD–metabolism discordance reported previously largely reflects statistical uncertainty in $CMRO_2$ estimates rather than widespread physiological sign reversal.

Introduction

In a recently published study, Epp et al. (2025) report widespread voxel-wise discordance between functional magnetic resonance imaging (fMRI) based blood-oxygenation-level-dependent (BOLD) signal changes and quantitative estimates of cerebral oxygen metabolism, concluding that BOLD responses frequently reflect metabolic changes of the opposite sign to those predicted by canonical neurovascular coupling. The findings have been widely covered by news outlets and have prompted broad discussion within the field about the physiological interpretability of BOLD fMRI and its implications for prior and ongoing neuroimaging work.

Extensive empirical work has examined how BOLD signal changes relate to underlying neuronal and metabolic processes. Simultaneous electrophysiology-fMRI experiments in non-human primates demonstrated that BOLD amplitudes covary with local field potentials (Logothetis et al., 2001), and gamma-band activity was shown to be linked to local hemodynamic signals and to

predict BOLD signal variance (Magri et al., 2012; Niessing et al., 2005). However, it is also established that the origin and interpretability of BOLD signal is complex, dynamic and far from understood (see for reviews: Buxton, 2012; Kim & Ogawa, 2012; Logothetis, 2008). In particular, negative BOLD responses remain difficult to interpret, as signal reductions can have heterogeneous metabolic underpinning (Godbersen et al., 2023; Stiernman et al., 2021). Whereas most of earlier work emphasized the need for caution in interpreting BOLD signal changes, Epp et al. suggest that the implications may be more substantial, reporting that the canonical interpretation of BOLD signal changes may lead to misinterpretation in approximately 40% of reported BOLD responses.

Epp et al. defined responses as *discordant* when they found opposing signs between task-evoked changes in BOLD (ΔBOLD) and quantitative cerebral metabolic rate of oxygen ($\Delta CMRO_2$). $\Delta CMRO_2$ was estimated using a biophysical model of oxygen delivery and extraction (He & Yablonskiy, 2007). The model combines arterial spin labeling–based cerebral blood flow (CBF) with venous deoxyhaemoglobin estimates derived from BOLD and R2' relaxometry to infer changes in oxygen consumption. This approach provides an indirect fMRI-based alternative to oxygen-tracer positron emission tomography (PET), which is considered the gold standard for $CMRO_2$ measurement (Ito et al., 2021). As with other quantitative fMRI approaches, $\Delta CMRO_2$ estimates depend on multiple model assumptions and are inherently noisier than PET-based metabolic measures (see for example Bright et al., 2019). Following this, one alternative possibility raised by Huth (2026) is that the reported discordance is due to measurement noise. Using simulations in which effect sizes were taken from the study and noise levels were set to reasonable, moderate values, Huth showed that even perfectly aligned underlying signals yield discordance rates of about ~40%, demonstrating that the observed discordance could arise as a consequence of noisy estimates rather than opposing physiology. To our surprise, Epp et al. did not explicitly account for noise or variability in general in $CMRO_2$ estimates, instead treating group-averaged values as a direct marker for metabolic changes, no matter how much single participant estimates varied. Building on this observation, and enabled by the open and transparent sharing of code and data by Epp et al., we reanalysed the original dataset to assess to what extent voxel-wise concordance or discordance classifications were supported by statistically robust $\Delta CMRO_2$ estimates.

Results and Discussion

The dataset comprises quantitative fMRI measurements of hemodynamic and metabolic parameters, including BOLD signal, cerebral blood flow (CBF), cerebral blood volume (CBV), effective transverse relaxation time (T2*), reversible transverse relaxation rate (R2'), and model-based estimates of cerebral metabolic rate of oxygen ($CMRO_2$). Data were acquired under two task conditions (calculation task and control, see Methods), and task-evoked changes were quantified as percent signal differences between conditions (ΔBOLD and $\Delta CMRO_2$). We first replicated the original concordance analysis by calculating voxel-wise sign mismatch between ΔBOLD and $\Delta CMRO_2$. As in Epp et al., we identified voxels showing statistically significant task-evoked BOLD responses (Fig. 2A; hereafter 'BOLD activation mask') and compared the signs of group-averaged ΔBOLD and $\Delta CMRO_2$ within this mask. Consistent with Epp et al., approximately 35% of voxels exhibited opposing signs of ΔBOLD and $\Delta CMRO_2$. However, this analysis assumes reliable and participant-consistent voxel-wise $\Delta CMRO_2$ estimates without formally evaluating either. Two examples illustrate this issue.

Example 1 (high variance)

A voxel shows a significant ΔBOLD of +0.5%. The group-mean $\Delta CMRO_2$ is −5%, and it is therefore labelled discordant. However, participant-level $\Delta CMRO_2$ estimates range from −50% to +40% with a high standard deviation. Such variability indicates that the group mean might not be a robust measure for the central tendency of the data in this case. Despite this, the voxel would still be classified as discordant.

Example 2 (inconsistent sign across participants)

Another voxel shows ΔBOLD of +0.5% and a small positive group-mean $\Delta CMRO_2$ of +0.2%, and is therefore labelled concordant. Yet 22 out of 40 participants show opposite signs of ΔBOLD and $\Delta CMRO_2$ (concordance rate = 0.45). Despite this discordant pattern on participant-level, the voxel is classified as concordant.

In both cases, classification is driven solely by the group mean, without evaluating whether the direction of $\Delta CMRO_2$ is statistically robust. Importantly, replacing the mean with the median (as done by Epp et al.) does not solve this issue, as neither statistic captures uncertainty or sign consistency. To quantify the issue of variability in $\Delta CMRO_2$ across participants, we examined variance descriptively across contrasts (Fig. 1). For each measure of interest (CBF, CBV, T2*, R2', and $CMRO_2$), we computed the coefficient of variation (CV) in each voxel as the standard deviation of the measure across subjects divided by the mean across subjects. This ratio indicates how consistent each measure is across participants, reflecting both inter-individual variability and noise. In both baseline and task conditions, $CMRO_2$ showed the highest CV. In percent signal change maps, BOLD exhibited a left-shifted and markedly narrower CV distribution, indicating lower relative variability across participants, whereas $CMRO_2$ showed both higher central CV values and substantially broader spread. Quantitatively, the median CV of $\Delta CMRO_2$ [= 4.72] was 26.1% higher than ΔBOLD [=3.74] across all cortical voxels and 127.8% [3.72 against 1.63] higher within BOLD-significant voxels, indicating markedly elevated inter-individual variability of metabolic estimates.

These observations demonstrate that $\Delta CMRO_2$ exhibits substantial variance, raising concerns about sign classification based solely on group averages. To address this limitation, we implemented two complementary inferential approaches. The first evaluated the stability of the sign relationship between ΔBOLD and $\Delta CMRO_2$ across participants. The second tested whether $\Delta CMRO_2$ differed significantly from zero and assigned concordant/discordant labels only for voxels showing a significant metabolic effect; all other voxels were categorized as showing no significant $\Delta CMRO_2$ change.

First, we tested whether the voxel-wise sign concordance rate across participants differed from chance. For each voxel, we computed the proportion of participants showing matching signs (positive or negative) for ΔBOLD and $\Delta CMRO_2$, and tested this proportion using a binomial test against a null hypothesis of 0.5, corresponding to chance-level agreement between the signs of ΔBOLD and $\Delta CMRO_2$. Voxels with concordance rates not significantly different from 0.5 cannot be robustly classified as either concordant or discordant. Across voxels, concordance rates clustered near 0.5 (mean = 0.54; Fig. 2B). Only 65 out of the 19,190 voxels in the BOLD activation mask (0.3%) were statistically classified as concordant and only 10 (0.05%) as discordant (FDR-corrected q = 0.05; empirical thresholds: $p_{concordant} \geq 0.81$, $p_{discordant} \leq 0.19$). Although this test is conservative, as it requires a strong and statistically reliable deviation from chance across participants, this analysis indicates that, for most voxels, the sign relationship between ΔBOLD and $\Delta CMRO_2$ is not consistent across participants and group-based classification should be interpreted with caution.

Second, we applied a group-level statistical criterion that likely provides a less conservative assessment of sign relationships than the participant-level concordance test described above. Specifically, we tested whether $\Delta CMRO_2$ differed significantly from zero across participants and classified voxels as concordant (matching signs of significant ΔBOLD and significant $\Delta CMRO_2$) or discordant (opposing signs) when $\Delta CMRO_2$ reached statistical significance at the group level, and otherwise as showing no significant $\Delta CMRO_2$ effect (see Methods). Within the BOLD activation mask, 77.2% of voxels showed no statistically significant group-level $\Delta CMRO_2$ effect and therefore could not be robustly classified as concordant or discordant. Among the remaining voxels (22.8%), 17.4% of all voxels were classified as concordant and 5.4% as discordant (Fig. 3A–B).

Clear differences in sign concordance patterns were observed between positive and negative BOLD activation maps (Fig. 3B,C). Within the positive BOLD mask, discordance was rare (0.4%), whereas 26% of voxels were concordant and 73.6% showed no statistically significant $\Delta CMRO_2$. In

Figure 1. Kernel density distributions of voxel-wise coefficients of variation (CV = SD/|mean| across participants) for CBF, CBV, T2*, R2', $CMRO_2$, and BOLD.

Left and middle panels show CV distributions under baseline [control task] and task [calculation] conditions, respectively. Right panel shows CV distributions of percent signal change (PSC) estimates.

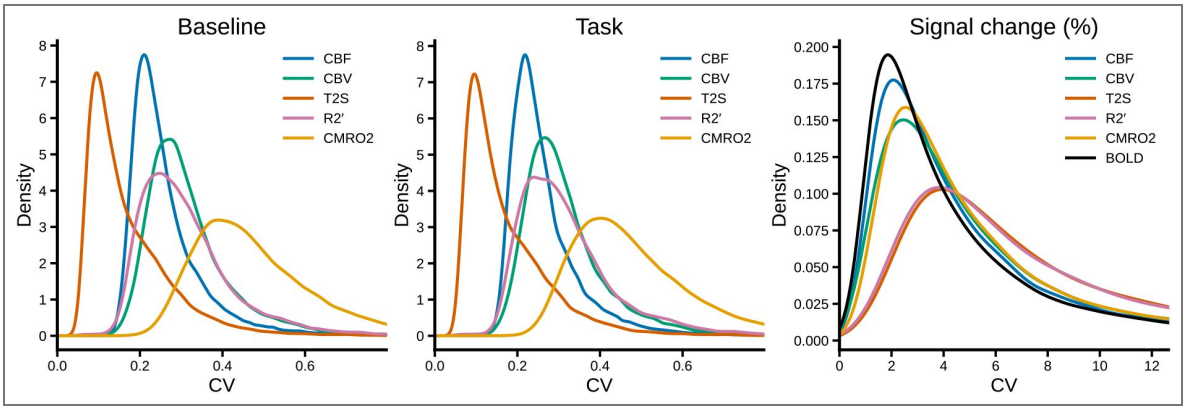
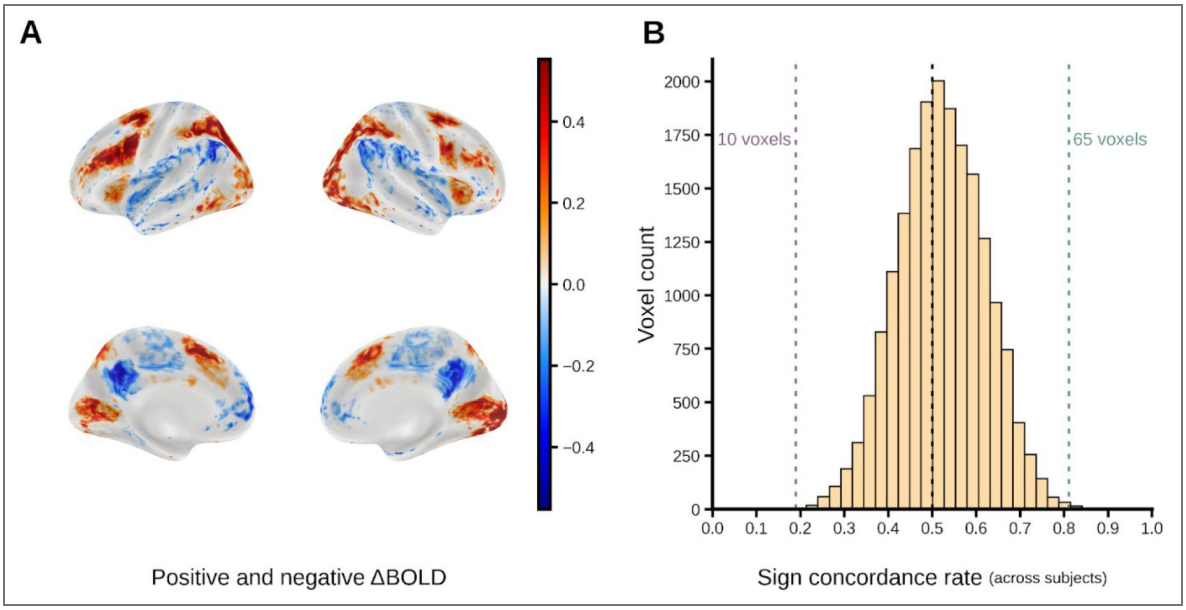


Figure 2.

A Cortical surface maps showing the mean change in BOLD signal (ΔBOLD) across participants for all voxels included in the BOLD activation mask. Warm colors (red–yellow) indicate positive ΔBOLD values, whereas cool colors (blue) indicate negative ΔBOLD values. The color bar represents mean ΔBOLD (percent signal change). **B** Histogram of voxel-wise sign concordance rates across participants within the BOLD activation mask. The concordance rate reflects the proportion of participants showing the same sign (positive or negative) of ΔBOLD and $\Delta CMRO_2$ at a given voxel. Vertical dashed lines denote q(FDR) <0.05 significance thresholds for classifying voxels as discordant (sign concordance below chance) or concordant (sign concordance above chance). The numbers above the dashed lines indicate the number of voxels falling beyond each threshold.



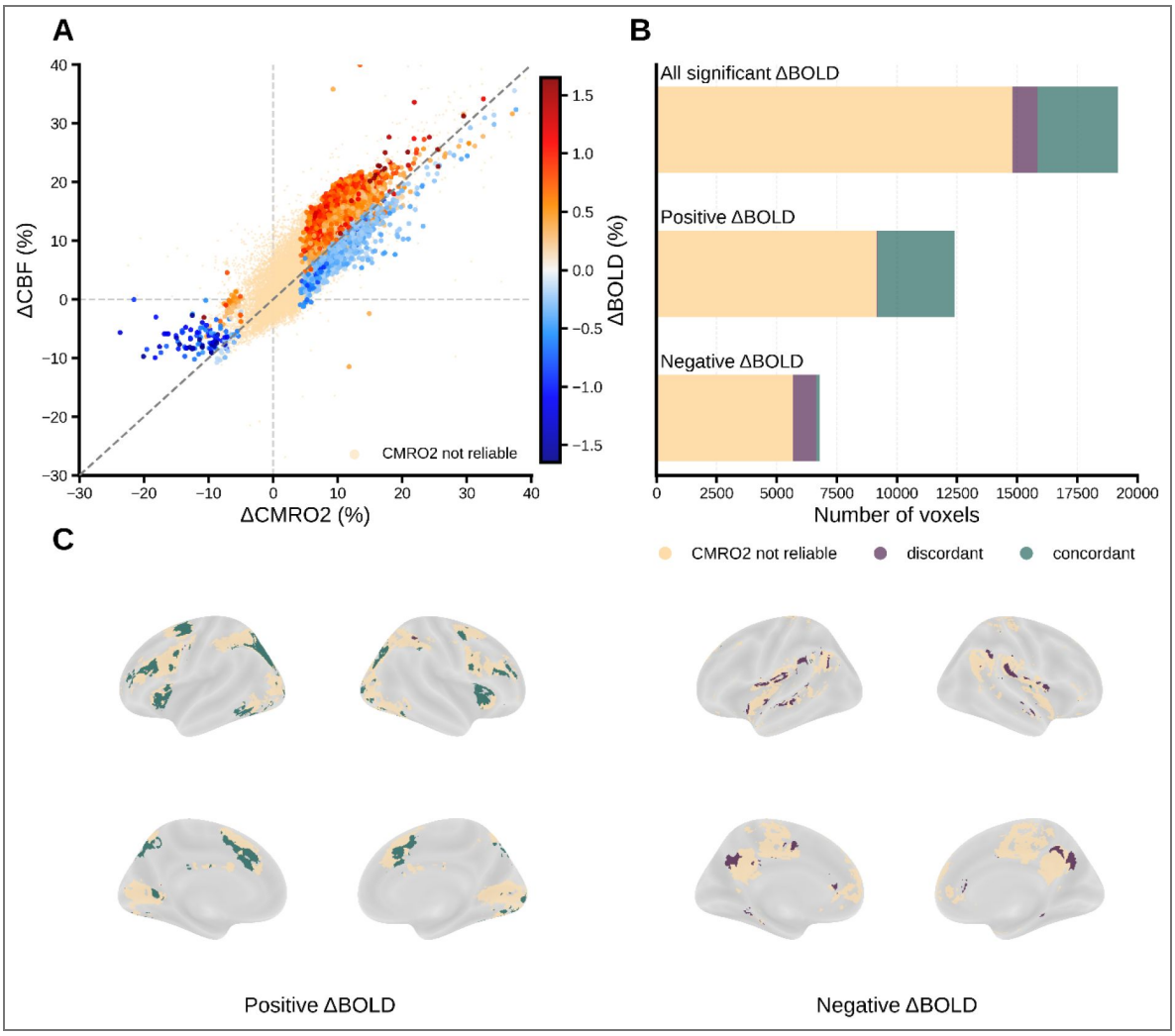


Figure 3.

A Relationship between median task-evoked changes in cerebral blood flow (ΔCBF) and cerebral metabolic rate of oxygen ($\Delta CMRO_2$) across all voxels showing statistically significant ΔBOLD responses. Voxels with statistically reliable $\Delta CMRO_2$ are color-coded by ΔBOLD sign and magnitude, whereas voxels with non-significant $\Delta CMRO_2$ are shown in light orange. The dashed diagonal indicates equal relative changes in CBF and $CMRO_2$; voxels above the diagonal for negative $\Delta CMRO_2$ and below the diagonal for positive $\Delta CMRO_2$ correspond to an n-ratio ($\Delta CBF/\Delta CMRO_2$) < 1. **B** Distribution of voxels within the BOLD activation mask by $CMRO_2$ reliability and sign concordance. The top bar shows all voxels with significant ΔBOLD responses (n = 19,190), partitioned into voxels with no significant $\Delta CMRO_2$ effect (77.2%), concordant voxels (17.4%), and discordant voxels (5.4%). The middle bar shows voxels with positive ΔBOLD, of which 26.0% were classified as concordant, 0.4% as discordant, and 73.6% showed no significant $\Delta CMRO_2$ effect. The bottom bar shows voxels with negative ΔBOLD, of which 1.7% were concordant, 14.6% discordant, and 83.7% showed no significant $\Delta CMRO_2$ effect. Bar lengths indicate the number of voxels in each category. **C** Cortical surface maps illustrating the spatial distribution of voxel classifications shown in (B). Left panels display voxels with positive ΔBOLD; right panels display voxels with negative ΔBOLD. Colors correspond to concordant, discordant, and $CMRO_2$ not reliable classifications.

contrast, within the negative BOLD mask, 14.6% of voxels were discordant and 1.7% concordant, while 83.7% showed no significant $\Delta CMRO_2$ effect. Crucially, these results indicate that, for most voxels classified as concordant or discordant by Epp et al., $\Delta CMRO_2$ estimates do not provide sufficient statistical support for robust sign determination. Where classification was possible, positive ΔBOLD was overwhelmingly concordant with metabolism, whereas negative ΔBOLD exhibited considerable higher rates of sign opposition, consistent with prior PET–fMRI findings (Godbersen et al., 2023; Stiernman et al., 2021).

Taken together, our analyses indicate that the central claim of Epp et al. – that approximately 40% of voxels exhibit metabolic responses opposite in sign to the observed BOLD signal – is not supported once the statistical robustness of $\Delta CMRO_2$ estimates is taken into account. Rather than indicating systematic physiological sign reversal, the reported voxel-wise discordance appears largely attributable to substantial variability in $\Delta CMRO_2$ estimates across participants, with 77.2% of voxels showing no statistically reliable direction of metabolic change. While some variability may reflect genuine inter-individual differences, it likely also reflects elevated noise levels for quantitative fMRI-based $CMRO_2$ estimation. When this variability is taken into account, the resulting pattern is largely consistent with prior literature: negative BOLD responses remain difficult to interpret due to heterogeneous underlying mechanisms, whereas for positive BOLD responses, where $\Delta CMRO_2$ effects reached statistical significance at the group level, metabolic changes were predominantly concordant with the observed BOLD signal. We therefore view the study by Epp et al. as an important contribution that stimulates discussion about BOLD interpretation, while suggesting that its conclusions should be considered in light of the statistical uncertainty of the underlying metabolic estimates.

Finally, we would like to commend Epp et al. for their commitment to open science principles. By making their data publicly available, they have enabled the kind of critical dialogue that is essential for scientific progress, rigor and transparency of the field.

Materials and Methods

Dataset and Participants

We reanalysed the openly available dataset used by Epp et al. (2025). Data were retrieved from OpenNeuro (*https://openneuro.org/datasets/ds004873*). Of the 40 participants included in the original study, we were able to retrieve complete data for 38 participants; two participants were excluded due to missing derivatives. All analyses were conducted on these 38 participants.

MRI data were acquired on a 3 T Philips Ingenia MRI scanner equipped with a 32-channel head coil. The quantitative fMRI protocol combined multiparametric quantitative BOLD (mqBOLD) imaging with arterial spin labeling (ASL) to estimate hemodynamic and metabolic parameters. Multiecho spin-echo T2 mapping was acquired using a 3D gradient spin-echo readout with eight echoes (TE_1 = ΔTE = 16 ms; TR = 251 ms; flip angle = 90°; voxel size = 2 × 2 × 3.3 mm^3; 35 slices). Multiecho gradient-echo T2* mapping was acquired with 12 echoes (TE_1 = ΔTE = 5 ms; TR = 2,229 ms; flip angle = 30°; voxel size = 2 × 2 × 3 mm^3; 35 slices). Cerebral blood volume (CBV) was measured using dynamic susceptibility contrast (DSC) MRI with single-shot GRE-EPI (TR = 2.0 s; flip angle = 60°; voxel size = 2 × 2 × 3.5 mm^3; 35 slices; 80 dynamics) following administration of a gadolinium-based contrast agent (0.1 ml kg^{-1}). Cerebral blood flow (CBF) was measured using pseudocontinuous arterial spin labeling (pCASL; post-labeling delay = 1,800 ms; labeling duration = 1,800 ms; TE = 11 ms; TR = 4,500 ms; voxel size = 3.28 × 3.5 × 6.0 mm^3; 20 slices; 39 dynamics plus M0 scan). Task-based BOLD fMRI was acquired using single-shot echo-planar imaging (TR = 1.2 s; TE = 30 ms; flip angle = 70°; voxel size = 3 × 3 × 3 mm^3; 40 slices; multiband factor = 2; 400 dynamics). A B0 field map was acquired for susceptibility correction (TR = 525 ms; TE_1/TE_2 = 6.0/9.8 ms; voxel size = 3 × 3 × 3 mm^3). These measurements were combined in the original study to derive voxel-wise maps of oxygen extraction fraction (OEF) and cerebral metabolic rate of oxygen consumption ($CMRO_2$) using the mqBOLD framework and Fick's principle. Full acquisition details are reported in Epp et al. (2025).

Task

Participants performed four task conditions in the original experiment: a calculation task (CALC), an autobiographical memory task, a low-level control condition (CTRL), and a resting-state baseline, as described in Epp et al. (2025). In the present reanalysis we focused exclusively on the CALC and CTRL conditions, as these were the main focus of Epp et al. as well.

In the CALC condition, participants solved arithmetic problems presented visually. Each trial displayed a row of three numbers followed by a question mark (n_1, n_2, n_3, ?), and participants were instructed to determine the missing number according to the arithmetic rule governing the sequence. Responses were selected from three answer options using the button box, with a maximum response time of 10 s per trial.

In the CTRL condition, participants performed a low-level baseline task with minimal cognitive demand. A row of random letters was presented for several seconds, and participants indicated via button press whether the first letter was a vowel. This condition was intended to provide visual input and motor responses comparable to the CALC task.

Preprocessing and Spatial Normalization

We did not perform preprocessing of the raw MRI data. Instead, we used the derivative datasets released by Epp et al. together with the original publication (*https://openneuro.org/datasets/ds004873*). The full preprocessing and modelling pipeline is described in detail in Epp et al. (2025); here we briefly summarize the relevant steps.

Preprocessing of BOLD fMRI data included motion correction, susceptibility distortion correction using field maps, as well as estimation and removal of confound regressors. Functional images were coregistered to the individual T1-weighted anatomical image using boundary-based registration implemented in FSL, and subsequently normalized to MNI152NLin6Asym standard space (2 mm isotropic resolution) using nonlinear registration implemented in ANTs 2.3.3.

Quantitative parameter maps used for metabolic modelling were computed in subject space from multiparametric qBOLD and ASL data using in-house scripts (MATLAB) and SPM12, as described by Epp et al. These included maps of T2, T2*, R2', cerebral blood flow (CBF), cerebral blood volume (CBV), oxygen extraction fraction (OEF), and cerebral metabolic rate of oxygen consumption ($CMRO_2$). All parameter maps were first coregistered to the first echo of the T2 acquisition and subsequently aligned to the participant's native T1-weighted anatomical image.

Spatial normalization to standard space was performed by applying the T1 to MNI nonlinear transformation estimated by fMRIPrep to all parameter maps. Where MNI-space derivatives were not directly available in the shared dataset, native-space parameter maps were transformed into MNI152 space using the transformation fields provided by the original authors.

All analyses were conducted in MNI152 space at 2 mm isotropic resolution using the group-level gray matter mask provided with the derivative dataset. To ensure consistency with the original study, we exclusively used the transformation fields and processing scripts released by Epp et al. (*https://github.com/NeuroenergeticsLab/two_modes_of_hemodynamics*) and did not introduce additional spatial preprocessing steps.

Percent Signal Change Maps

To ensure consistency with the original analysis, we used the percent signal change (PSC) maps provided in the derivative dataset released by Epp et al. (2025). These maps quantify task-evoked BOLD signal changes for the CALC condition relative to the CTRL baseline and are provided in MNI152 standard space (2 mm isotropic resolution).

PSC maps were derived from the preprocessed BOLD time series using the procedure implemented in the original preprocessing pipeline. For each participant and voxel, percent signal change was calculated as the relative difference between the BOLD signal during the CALC task

and the baseline signal during the CTRL condition, expressed as a percentage of the baseline signal. Voxels with invalid baseline values were excluded during PSC computation to avoid numerical instability.

Task-evoked changes in $CMRO_2$ were computed by comparing $CMRO_2$ estimates obtained for the CALC and CTRL conditions. In the original study, $CMRO_2$ maps were derived using a semiquantitative, BOLD-informed approach in which baseline quantitative parameter maps (R2', CBV, and CBF) were combined with task-related BOLD signal changes to estimate condition-specific $CMRO_2$ via the mqBOLD framework and Fick's principle. In this approach, task-related changes in R2' are approximated from baseline R2' values and the observed BOLD signal change, rather than being estimated directly from multiecho acquisitions during task conditions, with the aim of reducing noise propagation. Percent signal change maps for $CMRO_2$ were then calculated as the relative difference between CALC and CTRL $CMRO_2$ estimates.

Voxel Selection: BOLD Activation Mask

Similar to Epp et al. (2025), we restricted subsequent analyses to voxels showing statistically significant task-evoked BOLD responses at the group level. In the original study, significant task-related BOLD effects were identified using a multivariate partial least squares (PLS) analysis that extracted latent variables relating BOLD activity to task conditions. In the present analysis, we instead defined the BOLD activation mask using a voxel-wise univariate approach to enable straightforward statistical interpretation. In contrast to voxel-wise statistical testing, PLS evaluates significance at the level of latent variables, while voxel contributions are interpreted using bootstrap ratios rather than formal voxel-wise p-values. Consequently, thresholded bootstrap maps may contain large numbers of voxels exceeding the chosen threshold without providing direct voxel-wise significance inference. We therefore adopted a conventional voxel-wise statistical framework to define task-responsive voxels. For each voxel, a one-sample t-test against zero was performed across participants on the CALC–CTRL percent signal change values. Resulting p-values were corrected for multiple comparisons using Benjamini–Hochberg false discovery rate (FDR) correction at $q = 0.05$. Voxels surviving correction defined the BOLD activation mask. Positive and negative ΔBOLD voxels were considered separately in subsequent analyses where indicated.

Coefficient of Variation Analysis

To quantify inter-individual variability in measured parameters, we computed voxel-wise coefficients of variation (CV) across participants. The CV was defined as the ratio of the standard deviation to the absolute value of the mean across participants ($CV = SD/|mean|$) and was calculated separately for each voxel in MNI space.

CV values were computed for baseline (CTRL) and task (CALC) conditions using the corresponding parameter maps provided in the derivative dataset. Specifically, CV distributions were evaluated for CBF, CBV, T2*, R2', and $CMRO_2$. For each parameter and condition, voxel-wise CV values were calculated across participants and subsequently aggregated across cortical voxels to obtain distributional summaries.

In addition, CV values were computed for PSC maps derived from the CALC–CTRL contrast. PSC maps were available for BOLD and the quantitative parameters used in the mqBOLD framework, allowing direct comparison of relative variability in task-evoked signal changes across modalities.

For visualization, voxel-wise CV values were pooled across voxels and plotted as kernel density distributions for each parameter. Separate distributions were generated for baseline (CTRL), task (CALC), and PSC maps (Fig. 1). These distributions were used to compare the relative variability of physiological parameters across participants under baseline conditions, during task performance, and for task-evoked signal changes.

As CV depends on the magnitude of the mean, CV values may be inflated in regions where mean signal changes are close to zero. The comparison therefore primarily serves as a descriptive indicator of relative variability rather than a formal statistical test.

Voxel-wise Sign Discordance (Replication Analysis)

To replicate the primary analysis of Epp et al., we computed voxel-wise sign discordance between group-mean ΔBOLD and group-mean $\Delta CMRO_2$ within the BOLD activation mask. Voxels were classified as concordant when the signs of the group-averaged ΔBOLD and $\Delta CMRO_2$ matched, and as discordant when the two signals exhibited opposing signs.

Participant-level Sign Concordance Analysis

To assess the robustness of voxel-wise sign classification across participants, we computed, for each voxel, the proportion of participants showing concordant sign between ΔBOLD and $\Delta CMRO_2$. A concordance rate of 0.5 indicates equal concordance and discordance across participants. We tested whether the observed concordance proportion differed significantly from 0.5 using a binomial test at each voxel. Resulting p-values were corrected for multiple comparisons using FDR ($q = 0.05$). Voxels were classified as:

1. Concordant: significantly greater concordance than 0.5,
2. Discordant: significantly less concordance than 0.5,
3. Indeterminate: not significantly different from 0.5.

Group-level $\Delta CMRO_2$ Classification

As a complementary and typically less conservative approach, we evaluated whether $\Delta CMRO_2$ differed significantly from zero at the group level using a one-sample t-test across participants. Resulting p-values were corrected for multiple comparisons using FDR ($q = 0.05$). This approach is typically less conservative than the participant-level concordance test because it evaluates whether the group mean differs from zero and does not require a supermajority of participants to exhibit the same sign relationship.

Within the BOLD activation mask, voxels were categorized as follows:

1. Concordant: $\Delta CMRO_2$ significantly different from zero and matching the sign of ΔBOLD,
2. Discordant: $\Delta CMRO_2$ significantly different from zero and opposite in sign to ΔBOLD,
3. No significant $\Delta CMRO_2$: $\Delta CMRO_2$ not significantly different from zero.

This classification allowed us to determine in how many voxels the direction of metabolic change could be inferred with statistical support.

Statistical Thresholds and Software

All statistical analyses were conducted in Python using NumPy, SciPy and Statsmodels. One-sample t-tests (scipy.stats.ttest_1samp) were used for group-level mean testing. Binomial tests were used for participant-level sign consistency. Multiple comparisons were controlled using Benjamini–Hochberg FDR correction with $q = 0.05$. Figures were generated using Matplotlib and Nilearn for surface visualization.

Data availability

The original dataset generated by Epp et al. is publicly available via OpenNeuro (*https://openneuro.org/datasets/ds004873*). All scripts used for data processing, statistical analyses, and figure generation, as well as summary statistics associated with this manuscript, are available in a public GitHub repo: *https://github.com/olegolt/BOLD_metabolism_reanalysis*

Additional information

Funding

Funder	Grant reference number	Author
EC | European Research Council (ERC)	https://doi.org/10.3030/883892	Ole Goltermann Christian Büchel

Author ORCID iDs

Alexander Huth: https://orcid.org/0000-0002-5031-5348
Christian Büchel: https://orcid.org/0000-0003-1965-906X

References

Bright M. G., Croal P. L., Blockley N. P., Bulte D. P. (2019) Multiparametric measurement of cerebral physiology using calibrated fMRI. *NeuroImage, Physiological and Quantitative MRI* **187**:128-144 https://doi.org/10.1016/j.neuroimage.2017.12.049 | PubMed

Buxton R. B. (2012) Dynamic models of BOLD contrast. *NeuroImage, 20 YEARS OF fMRI* **62**:953-961 https://doi.org/10.1016/j.neuroimage.2012.01.012 | PubMed

Epp S. M., Castrillón G., Yuan B., Andrews-Hanna J., Preibisch C., Riedl V. (2025) BOLD signal changes can oppose oxygen metabolism across the human cortex. *Nature Neuroscience* **29**:1225-1236 https://doi.org/10.1038/s41593-025-02132-9 | PubMed

Godbersen G. M., Klug S., Wadsak W., Pichler V., Raitanen J., Rieckmann A., Stiernman L., Cocchi L., Breakspear M., Hacker M., *et al.* (2023) Task-evoked metabolic demands of the posteromedial default mode network are shaped by dorsal attention and frontoparietal control networks. *eLife* **12**:e84683 https://doi.org/10.7554/eLife.84683 | PubMed

He X., Yablonskiy D. A. (2007) Quantitative BOLD: Mapping of human cerebral deoxygenated blood volume and oxygen extraction fraction: default state. *Magnetic Resonance in Medicine* **57**:115-126 https://doi.org/10.1002/mrm.21108 | PubMed

Huth A. (2026) bold-cmro2-cbf-r2.ipynb [Jupyter notebook]. GitHub. https://github.com/alexhuth/notebook-sharing/blob/master/bold-cmro2-cbf-r2.ipynb

Ito H., Kubo H., Takahashi K., Nishijima K.-I., Ukon N., Nemoto A., Sugawara S., Yamakuni R., Ibaraki M., Ishii S. (2021) Integrated PET/MRI scanner with oxygen-15 labeled gases for quantification of cerebral blood flow, cerebral blood volume, cerebral oxygen extraction fraction and cerebral metabolic rate of oxygen. *Annals of Nuclear Medicine* **35**:421-428 https://doi.org/10.1007/s12149-021-01578-8 | PubMed

Kim S.-G., Ogawa S. (2012) Biophysical and Physiological Origins of Blood Oxygenation Level-Dependent fMRI Signals. *Journal of Cerebral Blood Flow & Metabolism* **32**:1188-1206 https://doi.org/10.1038/jcbfm.2012.23 | PubMed

Logothetis N. K. (2008) What we can do and what we cannot do with fMRI. *Nature* **453**:869-878 https://doi.org/10.1038/nature06976 | PubMed

Logothetis N. K., Pauls J., Augath M., Trinath T., Oeltermann A. (2001) Neurophysiological investigation of the basis of the fMRI signal. *Nature* **412**:150-157 https://doi.org/10.1038/35084005 | PubMed

Magri C., Schridde U., Murayama Y., Panzeri S., Logothetis N. K. (2012) The Amplitude and Timing of the BOLD Signal Reflects the Relationship between Local Field Potential Power at Different Frequencies. *Journal of Neuroscience* **32**:1395-1407 https://doi.org/10.1523/JNEUROSCI.3985-11.2012 | PubMed

Niessing J., Ebisch B., Schmidt K. E., Niessing M., Singer W., Galuske R. A. W. (2005) Hemodynamic Signals Correlate Tightly with Synchronized Gamma Oscillations. *Science* **309**:948-951 https://doi.org/10.1126/science.1110948 | PubMed

Stiernman L. J., Grill F., Hahn A., Rischka L., Lanzenberger R., Panes Lundmark V., Riklund K., Axelsson J., Rieckmann A. (2021) Dissociations between glucose metabolism and blood oxygenation in the human default mode network revealed by simultaneous PET-fMRI. *Proceedings of the National Academy of Sciences* **118**:e2021913118 https://doi.org/10.1073/pnas.2021913118 | PubMed

Samira Epp, Gabriel Castrillon, Beija Yuan, Jessica Andrews-Hanna, Christine Preibisch, Valentin Riedl (2025) BOLD signal changes can oppose oxygen metabolism across the human cortex. OpenNeuro. https://doi.org/10.18112/openneuro.ds004873.v2.0.6

Peer reviews

Reviewer #1 (Public review):

The study by Epp et al. has indeed gotten a lot of attention. As so often in the fMRI literature, some voices had taken the results out of proportion as if this result would suggest that we cannot trust fMRI. This is so, while informed researchers are aware of the capabilities and challenges of BOLD as a measure of neural activity. The paper was discussed and criticized on many aspects from various angles. E.g. with respect to unestablished models of estimating CMRO2, the 40% figure is being overestimated by the mask definition, and expected neuronal and vascular effects underlying the discordance.

The first publications of these discussions are being shared now. E.g. Chen et al. *https://doi.org/10.1038/s41593-026-02288-y*. The manuscript at hand augments this discussion. Specifically, the manuscript provides a direct statistical refutation of the recently proposed widespread physiological sign reversal between BOLD and CMRO2.

By reanalyzing a high-profile dataset, the authors demonstrate that the previously reported 40% discordance rate is an artifact of statistical uncertainty rather than a genuine physiological phenomenon. This critical re-evaluation restores some confidence in the canonical interpretation of BOLD signals that was recently challenged. It highlights the necessity of rigorous statistical validation in quantitative fMRI.

The following points should be addressed:

(1) Absence of evidence is taken as evidence of absence

The group-level significance analysis, summarized in the horizontal bar chart and cortical surface maps, labels non-significant voxels as 'CMRO2 not reliable', and the discussion concludes that positive BOLD responses are predominantly concordant with metabolism.

The paper treats voxels with non-significant CMRO2 effects as 'statistically uncertain' rather than as potentially reflecting genuine null metabolic changes, conflating absence of evidence with evidence of absence. Because the 77.2% of voxels shown as light orange could reflect either real null metabolism or insufficient power, the paper cannot distinguish between these. This ambiguity matters because a genuine null metabolic response to positive BOLD would itself be physiologically interesting and would not straightforwardly support 'predominant concordance'.

(2) Contextualization in other current literature

I feel that the introduction of the paper could also consider the embedding of the current literature about biophysical processes in the negative areas.

The negative responses have partly been discussed in the literature on quantitative physiology: e.g., Bohraus et al have been able to pinpoint the source of negative CMRO2 in positively activated voxels to large veins (*https://doi.org/10.1016/j.celrep.2023.113341*). Huber et al. have found that the neurovascular coupling (arterial venous weighting) is

different in positively and negatively activated brain areas, making the interpretation of derived parameters on physiology hard.

(3) Stylistic comments.

In places, the tone of the language could be revised to ensure that it is perceived as making a constructive contribution to the discussion.

https://doi.org/10.7554/eLife.111743.1.sa1

Reviewer #2 (Public review):

Summary:

The rebuttal aims to provide a statistical re-evaluation of Epp et al. to investigate the effects of CMRO2 uncertainty on concordance/discordance analysis between BOLD signal responses and CMRO2 change estimates based on an R2 framework. The authors observe markedly higher variance in CMRO2 compared to BOLD, which raises concerns about sign classification purely based on group means/medians.

Strengths:

The study is well motivated, and the analytical pipeline is rigorous and has been provided. Overall, the manuscript provides several thoughtful and rigorous analyses that contribute meaningfully to the ongoing discussion surrounding neurovascular coupling and $CMRO_2$ estimation.

Weaknesses:

Some aspects of the analytical framework could be improved, as well as the discussion of the caveats of the methods of this and the original paper.

(1) The binomial framework discussed on line 110 and described on line 321 reduces continuous ΔBOLD and ΔCMRO2 measurements to binary concordant/discordant labels, which may overemphasize unstable sign flips near zero effect sizes while discarding potentially meaningful magnitude information. The authors acknowledge that this overly strict approach yields very few meaningful voxels. A better justification or explanation of what we are meant to take away from this, other than the variability in the measurement, which is also explored elsewhere, would be helpful to the reader.

(2) In the methods, in the section entitled: Voxel Selection: BOLD Activation Mask, the authors describe their more traditional univariate statistical method as compared to the PLS approach used in the Epp paper. While I appreciate why the authors chose this approach, which simplifies interpretation, is it possible that this led to a lower number of discordant voxels? If yes, then I would suggest this be also added in the discussion of how the original Epp paper's methodological choices led to the very large percentage of discordant voxels.

(3) In the original paper, it looks to me like the discordant voxels have low CBF change and low rOEF. The gadolinium-based CBV measurement used to calculate OEF is a measure of total blood volume, while the blood volume that contributes to BOLD resides predominantly in veins and capillaries. Given the long PLD of the ASL acquisition and the total blood volume measurement, it seems to me that it is possible that discordant voxels may have high arterial blood volume, leading to overly large CBV measurement and an underestimation of CBF at this PLD (especially given their young age, for which I would expect ATT to be closer to 1-1.5s based on recent literature). While this is not currently discussed in this paper, it might be relevant to discuss how acquisition choices could bias some voxels towards erroneous CMRO2 estimates, which in turn would lead to these voxels being identified as discordant.

(4) In the methods, on line 267, the authors describe how they calculated ΔCMRO2 and how it differs from the original paper. A short discussion of how this choice is likely to affect the variance estimates would be warranted, given that the original paper seems to have chosen their method for the explicit purpose of decreasing error propagation. Especially, I wonder if this difference could account for the observation that "77.2% of voxels showed no statistically significant group-level $\Delta CMRO_2$ effect".

https://doi.org/10.7554/eLife.111743.1.sa0