Elsevier Editorial System(tm) for NeuroImage
Manuscript Draft

Manuscript Number: NIMG-11-633R2

Title: PHYCAA: Data-Driven Measurement and Removal of Physiological Noise in BOLD fMRI

Article Type: Regular Article

Section/Category: Methods & Modelling

Corresponding Author: Mr. Nathan W Churchill,

Corresponding Author's Institution: University of Toronto

First Author: Nathan W Churchill

Order of Authors: Nathan W Churchill; Grigori Yourganov; Robyn Spring; Peter M Rasmussen; Wayne Lee; Jon E Ween, M.D.; Stephen C Strother, Ph.D.

Abstract: The effects of physiological noise may significantly limit the reproducibility and accuracy of BOLD fMRI. However, physiological noise evidences a complex, undersampled temporal structure and is often non-orthogonal relative to the neuronally-linked BOLD response, which presents a significant challenge for identifying and removing such artifact. This paper presents a multivariate, data-driven method for the characterization and removal of physiological noise in fMRI data, termed PHYCAA (PHYsiological correction using Canonical Autocorrelation Analysis). The method identifies high frequency, autocorrelated physiological noise sources with reproducible spatial structure, using an adaptation of Canonical Correlation Analysis performed in a split-half resampling framework. The technique is able to identify physiological effects with vascular-linked spatial structure, and an intrinsic dimensionality that is task- and subject-dependent. We also demonstrate that increasing dimensionality of such physiological noise is correlated with increasing variability in externally-measured respiratory and cardiac processes. Using PHYCAA as a denoising technique significantly improves simulated signal detection with physiological noise, and real data-driven model prediction and reproducibility, for both block and event-related task designs. This is demonstrated compared to no physiological noise correction, and to the widely used RETROICOR (Glover et al., 2000) physiological denoising algorithm, which uses externally measured cardiac and respiration signals.

1. Cover Letter

Dear editor,

We have submitted the manuscript for the research article entitled “Data-Driven Measurement and Removal of Physiological Noise in BOLD fMRI”, by authors Nathan Churchill, Grigori Yourganov, Robyn Spring, Peter Rasmussen, Wayne Lee, Jon Ween and Stephen Strother. In this article, we propose a multivariate data-driven technique to characterize and remove vascular-linked physiological artifact in fMRI; this method identifies physiological noise based on temporal autocorrelation and spatial reproducibility, and requires no external measurements of cardiac or respiratory rates. We demonstrate that the algorithm significantly improves the prediction and reproducibility of fMRI results. In addition, the measured physiological noise has an intrinsic dimensionality that is both subject- and task-specific, and is correlated with variability in respiration and heartbeat. This paper not only presents an effective method for controlling physiological noise, but also provides a novel characterization of how experimental factors influence the physiological noise structure, which may better inform standard denoising methods.

Thank you for your consideration.

Sincerely,

Nathan Churchill

University of Toronto, Department of Medical Biophysics

***Highlights**

Research Highlights: > We propose multivariate, data-driven method for removing physiological noise in fMRI. > Identifies artifact via autocorrelation, power spectrum and spatial reproducibility. > Significantly improves prediction and reproducibility of fMRI results. > Noise shows subject/task-dependent dimensionality. > Noise dimensionality correlated with respiratory and cardiac rate variability.

***6. Response to Reviews**

General Comment.

We again thank the reviewers for their helpful, insightful comments; the advice has allowed us to significantly improve the quality of the submitted manuscript. We have addressed each comment in the sections below; please note that all major changes in the manuscript are underlined in-text.

Reviewers' comments:

Reviewer #1: The revisions made by the authors are generally excellent, and the elucidation of issues was both helpful and greatly appreciated. I am now strongly convinced of the efficacy of the approach, and look forward to implementing this in future analyses.

Comment 1: The only point that I believe still requires some clarification on the part of the authors is the description of Figure 3. I come away from the official caption and the first text paragraph describing this figure with the perception that I understand the material, however the second text paragraph (beginning on p. 18, "The spectral plots...") leaves me confused. In particular, this second paragraph refers to greater high-frequency power being associated with weaker autocorrelation, and then refers to this as being indicated by the blue-green to yellow pixels. However, in the preceding paragraph AND in the caption, the colorscale is (obliquely) indicated to refer to the fraction of energy in the power spectrum above 0.1 Hz (referred to as "high-frequency" in the statement under evaluation). Given that these colors refer to middle-range values in the colorbar (and some of the LOWEST in the figure), this description seems incongruent. Based on the caption, these colors refer to LOWER high-frequency power..? Further, I am unclear as to how "weaker autocorrelation" fits into this scheme. If the CVs are ranked/ordered from top-to-bottom by decreasing autocorrelation, it would seem that this upper CVs actually represent STRONGER autocorrelation. This may simply be a case of several terms being somewhat overloaded and colliding here, but the authors should re-evaluate the interplay between these two paragraphs and the figure for clarity.

I am also a bit unclear as to what the authors mean by "sufficiently high-frequency" in the second sentence for this same paragraph...I think they are simply trying to convey that there are detectable CVs that achieve the threshold of having 50% of their power above 0.1 Hz AND achieve significant autocorrelation (i.e., those above the black line...a simple statement that might be useful to state explicitly).

Response 1: we appreciate that the details of Fig. 3 are somewhat convoluted, although the reviewer is correct in their understanding of the results. We have rewritten the paragraph of interest, to further improve on the clarity of these results:

The spectral plots, displayed in Figs. 3(a,b), show similar structure for both tasks. Above the black line CVs are significantly autocorrelated ($p < 0.05$), and have a generally lower fraction of spectral power > 0.1 Hz. Below the black line, CVs are not significant ($p < 0.05$) and have a relatively high fraction of spectral power > 0.10 Hz. For PC dimensions < t/2 (40 PCs for TMT, 38PCs for SART), there are a subset of identifiable CVs that attain both greater than 50% of spectral power above 0.10 Hz, and significant autocorrelation (i.e. blue-green to yellow pixels, in a band above the black line). The number of significant CVs with more than 50% of power above 0.10 Hz tends to be relatively consistent across subjects, for a given PC dimensionality (Fig. 3(c,d)). However, for PC dimensions > t/2, a transition

gradually occurs, as mean spectral power is now observed to be more uniform across CVs, with increasing PC dimensionality. For these dimensionalities, an increasing proportion of components account for significant, high-frequency autocorrelation, rapidly increasing the possible number of identified physiological regressors (Fig. 3(c,d)).

Otherwise, from a picky perspective:

Comment 2-1. It would be helpful to include spaces in the 'nxt' notation (i.e., '\emph{n} x \emph{t}') where it exists...initially appears as if 'nxt matrix' is to be a meaningful term.

Response: spaces were placed between the letters.

Comment 2-2. On p. 11, a "respiratory cushion" is referred to, but a "belt" is referenced later in the same page --- same thing?

Response: this is correct, but we have replaced “respiratory cushion” with “pneumatic belt” for consistency.

Comment 2-3. On p. 15, "data was" should be "data were"

Response: error was corrected.

Comment 2-4. On p. 17, "...more than FWHM..." is noted here, but it would be more helpful to clarify which FWHM this refers to (filter? noise spatial correlation?)

Response: we have clarified that it is the FWHM of the spatial smoothing filter.

Comment 2-5. On p. 18, I'm not sure it is terribly useful to note that beyond t/2 "...where the majority of subject condition numbers...." I think this clause can be dropped without loss of clarity --- other changes that explain where these t/2 values come from now make the primary statement clear.

Response: agreed. We have removed this statement.

Reviewer #3: The authors present a response to reviews of a prior submission of a manuscript presenting a data-driven physiolgic noise correction method based on serially applied PCA and CCA. The authors have been largely responsive to the prior critiques, there are some additional major issues that need to be addressed before the manuscript is suitable for publication in NeuroImage.

Major comments:

Comment 1. Although the prior literature on data-driven methods that address physiologic noise in some ways is extensive, the authors failed to refer to particularly relevant data-driven physiologic noise removal methods that have been published in the past few years: Behzadi (CompCor) NI 37:90-101 and Beall (PESTICA) NI 37:1286-1300. Both of these prior methods have goals identical to this work and should be discussed. In particulary, the PESTICA method has gained some acceptance within the field and is freely

available from NITRC (www.nitrc.org/projects/pestica). This method has proven efficacious at both resting-state and activation-based fMRI data.

Response: we agree that it is important to validate this procedure against similar data-driven estimation models, in order to establish the optimal denoising approaches based on quantitative criteria. Comparing the suggested methods to PHYCAA would require greatly expanding the scope of the present paper, which already characterizes the algorithm, explores physiological noise structure, and provides validation against the standard RETROICOR approach. However, we have already begun evaluating a set of ICA-based methods in the same task-design framework; we have recently submitted a manuscript that demonstrates that PESTICA is of comparable importance to RETROICOR for denoising, with the former optimizing prediction and reproducibility metrics for 10/24 subjects, the latter optimizing 11/24 subjects, and 4/24 subjects requiring both preprocessing steps (see attached abstract, APPENDIX). Following the publication of PHYCAA results, we intend to perform a more extensive comparison of such procedures that includes PHYCAA. We agree that resting-state data will be of particular importance for these subsequent comparisons, but the rough equivalence between RETROICOR and PESTICA found in our submitted manuscript is for a subtle cognitive contrast (i.e. Trail Making B vs Trail Making A), which may not be too different from resting state studies. However, this remains to be investigated and is beyond the scope of this manuscript.

We have added a section in Discussion to address this issue for the current manuscript, and discuss the potential tradeoffs between a few of these methods, and motivation for the use of PHYCAA (p. 28, paragraph 2):

In recent years, a number of alternate data-driven models have also been developed, with the goal of estimating and removing physiological noise (Behzadi et al., 2007; Perlbarg *et al*., 2007; Tohka *et al*., 2008; Beall et al., 2010a,b). We suggest that the proposed PHYCAA model provides advantages in that it is built on a set of statistically rigorous, data-driven methods, requiring no qualitative estimation of spatial priors or spectral peaks. In addition, we have shown that identification of physiological noise requires adaptive subspace estimation, and that the reproducibility measures used can optimize subspace estimation of correlated, spatially-distributed BOLD signals (Yourganov et al., 2011). To our knowledge, no pre-existing method optimizes physiological noise estimation in such a quantitative resampling framework. By comparison, other techniques often perform estimation without using power-spectrum constraints (e.g. Behzadi et al., 2007; Perlbarg *et al*., 2007; Tohka *et al*., 2008), and certain cases, take advantage of lags in slice acquisition for increased temporal sensitivity (Beall et al., 2010a). It remains to be investigated which of these features of the various physiological noise estimators are of greatest individual and combined importance.

Comment 2. To-date, it is far more common and important to apply physiologic noise correction to resting state data, than to activation-based fMRI, particularly because the analysis methods used (e.g. correlation analysis) are more susceptible to contamination from the global spatiotemporal correlations caused by physiologic noise. It would be interesting to see a comparison of the impact of the proposed method on resting state data, but at the very least the application should be discussed.

Response: We agree that correcting for physiological noise in resting-state data is indeed of particular importance, given the sensitivity of connectivity and correlation analyses to structured artifacts. This may provide an effective area in which to further validate PHYCAA, for example testing against other popular

noise estimators, such as PESTICA and CompCor techniques. However, we note that it is unclear that standard resting state correlation-based analyses are more sensitive to preprocessing than the variance-driven, adaptive multivariate analyses used in our paper.

We have focused on fMRI data with an overt task design, partly because it allows for both predictive and reproducible model validation, but also to understand how physiological noise may be modulated by different types of cognitive engagement. However, the choice of tasks with continuous, asynchronous behaviour (e.g. the non task-locked tracing tasks of TMT, and GO-NOGO responses of SART) and weaker, more subtle contrast, allows us to infer that PHYCAA may have a significant denoising effect for resting state data as well.

We have amended Discussion (p. 29, paragraph 2) to address these points:

It is thus important to compare PHYCAA with preexisting data-driven models in a consistent resampling framework, to determine which methods optimize model prediction and reproducibility. Of particular interest is the denoising efficacy for resting-state analyses, which involve techniques that are often highly sensitive to the high-variance, high autocorrelation of physiological noise, and lack an overt task-design for predictive model validation (see Cole et al., 2010 for an overview of these issues). We have performed the current analyses using overt task designs in order to validate PHYCAA's effects, including predictive accuracy. However, in this paper we have examined two models that involve asynchronous, relatively weak cognitive contrasts, and used multivariate/locally-multivariate analysis models which are also sensitive to the high-variance effects of physiological noise, given their dependence on the underlying data covariance structure (Mardia et al., 1979; Strother et al., 2010). This provides some evidence of PHYCAA's potential efficacy in denoising, although direct comparison against other available techniques in a resting-state framework must be addressed in future work.

Comment 3. The principle that physiologic noise will be aliased into the higher frequencies (>0.1Hz), as stated by the authors, is counter-intuitive to this reviewer. In the data presented here, the TR is 2s, which means the nyquist frequency is 0.25Hz. In such dramatically undersampled data, cardiac rates in different subjects will be aliased arbitrarily throughout the sampled frequencies. I also don't see how the Glover or Windischberger references support this assumption. Please justify that cardiac effects will be aliased preferentially to sampled frequencies above 0.1Hz in acquisitions with TR>=2s, as was used here.

Response: The extensive aliasing of cardiac effects poses a difficult issue for separation from the BOLD signal, based on spectral power. We agree that the current citations do not provide sufficient rationale, and have added a section in Methods (p. 7, paragraph 1), which acknowledges that although cardiac-linked fluctuations with spectral power above 0.1 Hz are typically identifiable (as a range of spectral peaks are often observed in the temporally undersampled data), this only constitutes a subset of cardiac effects. While we are likely able to remove the majority of respiratory effects, the PHYCAA model can only remove a (subject dependent) portion of cardiac noise. We also note that the current model is insensitive to the lower-frequency aliased effects, and this is an area for further algorithmic improvements:

Physiological noise may also be characterized by power spectrum, as the effects of respiration and cardiac pulsation tend to be centered at frequencies of approximately 0.3 and 1.0 Hz, respectively, whereas BOLD oscillations are primarily localized in the 0.01-0.08 Hz spectral band (Zou et al., 2008;

Cole et al., 2010). For standard fMRI acquisition TRs > 1 s, respiration effects are often partly aliased, though spectral power is generally concentrated above 0.1 Hz (Glover et al., 2000), while cardiac effects tend to be more extensively aliased. However, the undersampled cardiac effects have been shown to be higher-order processes (Penny et al., 2003) that comprise multiple oscillatory "modes", producing multiple spectral peaks for a given dataset (e.g. Kiviniemi et al., 2003). Data-driven analyses consistently detect high-frequency pulsatile components in data, with characteristic vascular-coupled spatial structure (e.g. McKeown et al., 1998; Kiviniemi et al., 2003; Perlbarg et al., 2006; Behzadi et al., 2007). Zollei et al. (2003) have also shown that for the majority of subjects, CAA extracts components with spectral peaks above 0.08 Hz, that have both vascular-coupled spatial patterns and correlations with external cardiac rate measurements. We therefore expect that spectral power > 0.10 Hz is predominantly due to physiological noise (as per Biswal et al., 1995; Zou et al., 2008; Cole et al., 2010). Although a portion of cardiac artifact often aliases below 0.10 Hz (Lund *et al*., 2006), and physiological noise also produces intrinsically low-frequency fluctuations (Birn et al., 2006; Shmueli et al., 2009), we conservatively focus on extracting high-frequency components for the current model, which is separable from BOLD signal based on spectral priors.

Having established the effectiveness of this conservative model, we also discuss potential extensions that are less sensitive to this low-frequency limitation (Discussion; p. 28, paragraph 1).

As a final point, we have also produced post-hoc evidence that the current denoising model captures a significant component of cardiac effect in the presented results. Regarding the spatial maps of Fig. 6, maximum physiological Z-scores occur predominantly in the brainstem and large vessels, as opposed to the brain-edge and ventricle localization of respiratory effects (comparing with Fig.s 7-8 in Lund et al. 2006). In addition, noise dimensionality and cardiac variance are correlated (Fig. 5, although the coupling of cardiac and respiratory variance makes it difficult to isolate these effects).

Comment 4. A related point is the observation that the proposed method increases fMRI CNR, whereas RETROICOR, in some cases, will reduce functional signal. It is a fact that in some cases there will be overlap in the observed frequency of the paradigm and physiologic processes. In this case, the authors restriction to higher frequencies prevents removal of functional signal, whereas RETROICOR's phase matching will unavoidably remove some signal. That doesn't mean that the functional signal remaining in the PHYCAA-corrected data is not contaminated by physiologic noise. Please include a discussion of this in the text.

Response: We thank the reviewer for this excellent point, and agree that we had not directly established whether the removal of task-design components (and thus eliminating the task-frequency band) is driving the PHYCAA model performance.

To examine whether this is the case, we directly tested whether RETROICOR is biased by task correlations with respiratory and cardiac phases. To do so, we regressed out the task component from fMRI data (in the same manner as for PHYCAA) and used RETROICOR to estimate physiological artifact from the residual, which is then subtracted from the original data; we are thus certain not to be removing task-correlated BOLD signal from the data. The (*P*, *R*) of these results was compared against PHYCAA. In addition, we examined the correlation between undersampled cardiac/respiratory phase and task

responses, to determine the relative risk of task-correlation in the phases of the physiological processes. The methods are outlined (p. 18, paragraph 3), and results presented (p. 24, paragraph 3). In general, RETROICOR with the task effect removed offers no significant improvement in model (P,R), and respiratory/cardiac phases are generally uncorrelated with the task design. It is unclear why a set of subjects still become considerably worse with the application of the modified RETROICOR, even though we have ensured that we are not removing task-correlated signal; however, it is possible that RETROICOR is fitting and removing transiently task-linked effects that contribute to between-class discrimination. The findings are reviewed in Discussion (p. 25, paragraph 2).

The second issue, that we are restricting denoising to higher-frequency components, which leaves a risk of lower-frequency physiological noise, is an ongoing challenge. Our denoising model is currently a conservative one, for which we have clarified that only a portion of the physiological is controlled (Methods; p.7, paragraph 1) – although it still has a significant impact on model performance, and observes an intrinsic structure and dimensionality. We have also amended the Discussion (starting p.28, paragraph 1) to reflect this issue, as motivation for using more sophisticated selection methods, such as basing the spectral band on externally measured cardiac and respiratory rates, or avoiding this complication altogether and using complex phase data; this would enable us to resolve the issue of extracting lower-frequency physiological noise as well:

One potential limitation of the current technique is that we restrict noise component selection to higher frequencies > 0.1 Hz. As previously shown (Lund et al., 2006), cardiac phase effects may be extensively aliased into the task-frequency range; in addition, the effects of change in cardiac rate and respiratory volume may be present in the task-frequency spectral band (Birn *et al*., 2006; Shmueli *et al*., 2007; Chang et al., 2009). The current model is conservative, as it is presently designed to remove the more directly separable high-frequency portion of physiological noise. However, the PHYCAA framework is flexible in design, which may allow us to substitute criteria that are more sensitive to physiological noise across the full spectral range. For example, we may deliberately undersample external cardiac and respiratory measurements, to identify the expected spectral bands of heavily aliased physiological artifact in fMRI data, along with slower effects of change in respiratory depth and cardiac rate.

Comment 5. Point 4 raises the point that PHYCAA's superior performance to RETROICOR could be entirely due to the exclusion of the paradigm frequency from the regression process. This could easily be included in RETROICOR (i.e. exclude the paradigm frequency from RETROICOR's regression).

Response: see part-1 of the response to **Comment 4**.

APPENDIX: submitted abstract on ICA denoising in fMRI

Title:

Optimizing preprocessing and analysis pipelines for single-subject fMRI:

2. Interactions with Task Contrast

AUTHORS:

Churchill NW; Yourganov G; Oder A; Tam F; Graham SJ; Strother SC

ABSTRACT

A variety of preprocessing techniques are available to correct subject-dependant artifacts in fMRI, caused by head motion and physiological noise. Although it has been established that the chosen preprocessing steps (or "pipeline") may significantly affect fMRI results, it is not well understood how preprocessing choices interact with other parts of the fMRI experimental design. In this study, we examine how two experimental factors jointly interact with preprocessing: between-subject heterogeneity, and strength of the task contrast being examined. Relatively weak and strong levels of cognitive contrast were examined in an fMRI adaptation of the Trail-Making Test, with data from young, healthy adults. The relative importance of standard preprocessing with motion correction, physiological noise correction, motion parameter regression and temporal detrending were examined for the different task contrast strengths. We also tested two Independent Component Analysis (ICA) denoising procedures, including a model for manual removal of artifacts, and PESTICA (Beal and Lowe, J Neuro Sci Meth, 2010) based on physiological priors. Results were analyzed using Penalized Discriminant Analysis, and model performance measured via reproducibility and prediction metrics, in the data-driven NPAIRS framework. In addition, we applied simulation methods to examine potential biases in individual-subject optimization. Our results demonstrate that (1) individual pipeline optimization is not significantly more biased than fixed preprocessing pipelines. In addition, (2) when applying a fixed pipeline across all subjects, the type of task contrast significantly affects pipeline performance; in particular, the effects of different ICA models vary significantly by task contrast, and are not by themselves optimal preprocessing steps. Also, (3) selecting the optimal preprocessing pipeline for each subject improves model performance, with the weaker cognitive contrast being more sensitive to pipeline optimization, e.g. choice of ICA denoising approach. These results demonstrate that sensitivity of fMRI results is influenced not only by preprocessing choices, but by interactions with other components of the experimental design, particularly activation signal strength.

Submitted to: PLoS-ONE, 26/06/2011

Title:

PHYCAA: Data-Driven Measurement and Removal of Physiological Noise in BOLD fMRI

Authors:

Nathan W. Churchill[a,c]*, Grigori Yourganov[b,c], Robyn Spring[b,c], Peter M. Rasmussen[d,e], Wayne Lee[f], Jon E. Ween[g,h], Stephen C. Strother

Affiliations:

[a] Department of Medical Biophysics, University of Toronto. Toronto, Ontario, Canada

[b] Institute of Medical Science, University of Toronto. Toronto, Ontario, Canada

[c] Rotman Research Institute, Baycrest. Toronto, Ontario, Canada

[d] DTU Informatics, Technical University of Denmark, Kgs. Lyngby, Denmark

[e] The Danish National Research Foundation's Center for Functionally Integrative Neuroscience, Aarhus University Hospital, Denmark

[f] Diagnostic Imaging, Hospital for Sick Children, Toronto, Ontario, Canada

[g] Posluns Centre for Stroke and Cognition, Kunin-Lunenfeld Applied Research Unit, Baycrest, Toronto, Ontario, Canada

[h] Division of Neurology, Faculty of Medicine, University of Toronto, Toronto, Ontario, Canada

Corresponding Author:

Nathan W. Churchill
Rotman Research Institute, Baycrest
3560 Bathurst Street, Toronto ON, Canada
M6A 2E1
Email: nchurchill@rotman-baycrest.on.ca
Phone: (416) 785-2500 x3067
Fax: (416) 785-2862

ABSTRACT

The effects of physiological noise may significantly limit the reproducibility and accuracy of BOLD fMRI. However, physiological noise evidences a complex, undersampled temporal structure and is often non-orthogonal relative to the neuronally-linked BOLD response, which presents a significant challenge for identifying and removing such artifact. This paper presents a multivariate, data-driven method for the characterization and removal of physiological noise in fMRI data, termed PHYCAA (PHYsiological correction using Canonical Autocorrelation Analysis). The method identifies high frequency, autocorrelated physiological noise sources with reproducible spatial structure, using an adaptation of Canonical Correlation Analysis performed in a split-half resampling framework. The technique is able to identify physiological effects with vascular-linked spatial structure, and an intrinsic dimensionality that is task- and subject-dependent. We also demonstrate that increasing dimensionality of such physiological noise is correlated with increasing variability in externally-measured respiratory and cardiac processes. Using PHYCAA as a denoising technique significantly improves simulated signal detection with physiological noise, and real data-driven model prediction and reproducibility, for both block and event-related task designs. This is demonstrated compared to no physiological noise correction, and to the widely used RETROICOR (Glover *et al*., 2000) physiological denoising algorithm, which uses externally measured cardiac and respiration signals.

Keywords: BOLD fMRI, physiological noise, image processing, multivariate, data-driven

1. INTRODUCTION

One of the principal confounds in BOLD fMRI analyses is the presence of physiological noise. Respiration and pulsatile blood flow induce relatively large fluctuations in BOLD signal, of up to 10% and 40% from baseline respectively (Birn *et al*., 2006; Dagli *et al*., 1999), compared to the 1-5% signal change reflected in the haemodynamic response (Chen & Small, 2007). Such noise effects, which are quasi-periodic with the respiratory and cardiac cycles, are centered at frequencies of approximately 0.3 Hz and 1.0Hz, respectively (Glover *et al*., 2000). At standard EPI sampling rates of > 1 s, these time series are extensively aliased, and cause complex, highly autocorrelated temporal signal which is difficult to model and remove in fMRI data (Lund *et al*., 2006). In addition, physiological noise may correlate spatially and temporally with the BOLD response, creating a significant challenge in separating signal from noise. Given these confounds, it is common practise to model and remove physiological noise using denoising algorithms, in order to minimize physiological artifact. This is typically performed via either parametric modelling using external physiological measurements, or data-driven estimation of noise in fMRI data.

Parametric methods are often used to remove physiological noise after acquisition, by measuring the cardiac and respiratory cycles, and regressing out signal correlated with either phase (Glover *et al.*, 2000) or changes in frequency of the two processes (Birn *et al*., 2006; Shmueli *et al*., 2007; Chang *et al*., 2009). However, these methods are based on relatively strict model assumptions, such as a linear relation between cardiac/respiratory phase and BOLD effect (Glover *et al*., 2000), and a fixed parametric transfer function for modelling cardio-respiratory BOLD effect, consistent over all brain regions and subjects (Birn *et al*., 2006; Chang *et al*., 2009). In addition, such physiological noise correction may interact with other preprocessing choices, including motion correction and slice-timing correction (Jones *et al*., 2008; Churchill *et al*., 2010), requiring careful ordering of preprocessing steps. Denoising methods based on cardiac and respiratory phase also show mixed effects on the reproducibility and predictive accuracy of fMRI results (Churchill *et al*. in press). For these reasons, data-driven basis decomposition methods have shown increasing appeal as an alternative denoising method.

Data-driven models, used to estimate and separate BOLD signal from physiological noise, are typically based on decomposition methods, such as Principal Component Analysis (PCA) (Bullmore *et al*., 1996; Hansen *et al*., 1999; Tegeler *et al*., 1999; Thomas *et al*., 2002) and Independent Component

Analysis (ICA) (e.g. McKeown *et al.*, 1998; Beckmann *et al*., 2004; Perlbarg *et al*., 2007; Tohka *et al*., 2008). A variety of less common techniques have also been proposed, including estimating the dominant physiological frequency based on voxel-wise power spectra (Chuang *et al*., 2001), using wavelet bases (Hilton *et al.*, 1996; Jahanian *et al*., 2005) and clustering methods (Windischberger *et al*., 2003; Song *et al*., 2006). Such models attempt to identify spatially and/or temporally independent sources, from which BOLD signal and noise effects originate, retaining only the "signal" sources for further analyses.

In general, ICA-based methods are most commonly used for physiological noise estimation; however, such data-driven techniques are subject to *post-hoc* selection issues, as the identified sources must be classified into signal and noise categories. This is often performed via manual selection of ICs, which is subjective and rarely quantitatively validated (Kelly *et al*., 2010). More recently, quantitative methods have also been proposed, including maximizing intersession reproducibility of ICs (Wang & Peterson, 2008; Yang *et al.*, 2008), or IC selection based on characteristic spatiotemporal patterns (Perlbarg *et al*., 2007); machine learning algorithms have also been applied to identify artifactual ICs (De Martino *et al*., 2007). However, such methods suffer from additional complications, as physiological signal and noise sources are not necessarily separable with unconstrained PCA/ICA, and physiological noise artifacts are themselves spatially reproducible, being consistently localized near arterial tracts, densely vascularized grey matter (e.g. the visual cortex) and ventricles (Dagli *et al*., 1999; Birn *et al*., 2006).

As an alternative to the ICA techniques, this paper presents a multivariate data-driven method to identify and remove physiological noise, based on temporal autocorrelation properties and employing Canonical Correlation Analysis (CCA); this multivariate technique identifies patterns of maximal correlation between variables of two datasets. Molgedy & Schuster (1994) first used maximized autocorrelation as a constraint to identify biophysical signal sources, employing a neural-net model. This was extended to fMRI by Petersen *et al*., (2000), Lukic *et al*. (2002) and Friman *et al*. (2002), with all papers demonstrating the use of data-driven CCA to detect the BOLD haemodynamic response. They used CCA to identify temporal patterns in fMRI with maximum autocorrelation, afterwards selecting the subset of components that include the task-response. It was first shown by Zollei *et al*. (2003) that the CCA temporal autocorrelation method also produces spatial patterns that correspond with regions of high cerebral high vascular flow; they suggested that this method be used to spatially map large blood vessels.

This paper adapts the principles of maximizing temporal autocorrelation in a multivariate CCA framework, to demonstrate a method for removing physiological artifact, which we term PHYCAA (PHYsiological correction using Canonical Autocorrelation Analysis). This method identifies temporal patterns that are high-frequency, high-variance and significantly autocorrelated, and removes them via regression. The CCA dimensionality selection is performed by estimating reproducibility in a split-half resampling framework, similar to NPAIRS (Strother et al., 2002; 2010), which takes advantage of the spatially consistent location of such noise signals. This model presents an alternative to purely data-driven basis decomposition methods, by integrating a set of adaptive, quantitative physiological constraints into the source separation problem.

This denoising technique is tested on simulated, and different types of experimental data and analysis models, to show its general applicability. The PHYCAA algorithm is applied to (1) a simulated Gaussian network structure with physiological noise contamination from real subject measurements, (2) a block-design adaptation of the Trail-Making Test, analyzed using Penalized Discriminant Analysis (PDA), and (3) a fast event-related GO-NOGO task, analyzed with searchlight Gaussian Naïve Bayes (slGNB). The identified high-frequency physiological noise of fMRI has an intrinsic spatio-temporal structure and dimensionality that is primarily influenced by task design, but evidences significant inter-subject variability as well. Therefore, we directly link the PHYCAA algorithm with cardiac and respiratory processes, by showing that the dimensionality of the estimated physiological signal increases with increasing variability in cardiac and respiratory rates, and respiration depth. We then demonstrate that data regression based on this technique tends to significantly improve ROC signal detection in simulations, and in real data sets increases the predictive accuracy and spatial reproducibility of results for the different task designs and analysis models; these results improve on the commonly-used physiological correction of RETROICOR (Glover et al., 2000).

2. METHODS

We begin by discussing the underlying theory of Principal Component Analysis (PCA) and Canonical Correlation Analysis (CCA), which provides the basis for this denoising method; autocorrelation maximization is also defined in the CCA framework. Note that commonly used variables and

abbreviations are listed in Table I. The motivations for this denoising method are then outlined, along with the denoising algorithm. We then establish the synthetic and experimental datasets which are used to validate the denoising method, along with preprocessing and analysis methods. We examined (1) trends in CCA decomposition, as a function of PC dimensionality, and (2) the effects of maximizing spatial reproducibility in the estimated physiological noise data, as well as the relationship of this model with cardiac and respiratory cycles. Finally, (3) we evaluated the effects of the PHYCAA algorithm on both signal detection in simulated data, and prediction and reproducibility of experimental fMRI results, compared to standard RETROICOR preprocessing (Glover *et al*., 2000), as well as the effects of PHYCAA denoising on spatial activation maps.

2.1 Dimensionality Reduction on a PCA basis

Representing each fMRI image as an *n*-voxel vector, a set of *t* timepoints can be expressed as the *n* x *t* matrix ***X***. This matrix may be diagonalized by singular value decomposition, expressed as:

$$X = USV^T \qquad [1]$$

Where $U_{(n \times t)}$ is the spanning set of orthonormal image bases, $S_{(txt)}$ is a diagonal matrix of singular values, and $V_{(txt)}$ is the set of orthonormal timeseries bases. The datapoints (e.g. image volumes) of ***X*** may be projected into a reduced subspace spanned by $U_k = [u_1 \ldots u_k]$ $(k < t)$, by:

$$Q_k = U^T X = S_k V_k \qquad [2]$$

Giving coordinate matrix Q_k with reduced dimensionality *k*.

2.2 Canonical Autocorrelation Analysis

Canonical Correlation Analysis (CCA), first developed by Hotelling (1935), is used to estimate multivariate correlation between variables of two data sets. For random vectors ***X*** and ***Y*** (with *p* and *q* variables, respectively), CCA determines weight vectors *a* and *b* that give linear combinations:

$$c_x = a^T X \qquad c_y = b^T Y \qquad [3]$$

with maximized correlation, defined:

$$corr(c_x, c_y) = \frac{a^T S_{xy} b}{\sqrt{a^T S_{xx} a}\sqrt{b^T S_{yy} b}} \qquad [4]$$

where S_{xx} and S_{yy} are covariance matrices of $\boldsymbol{X}$ and $\boldsymbol{Y}$, and S_{xy} is the covariance between row-centered $\boldsymbol{X}$ and $\boldsymbol{Y}$. Note that for univariate $\boldsymbol{X}$ and $\boldsymbol{Y}$, this reduces to the Pearson correlation coefficient, and if $\boldsymbol{X}$ or $\boldsymbol{Y}$ is a categorical matrix of condition labels this is equivalent to Canonical Variates Analysis and Fisher's Linear Discriminant (CVA; Mardia *et al*., 1979). The model identifies K orthogonal pairs of c_x, c_y (where $K = \min(p,q)$), that maximize correlation ρ between variables of $\boldsymbol{X}$ and $\boldsymbol{Y}$, by solving the eigenvector decompositions:

$$(S_{xx}^{-1}S_{xy}S_{yy}^{-1}S_{yx})a = \rho^2 a \qquad (S_{yy}^{-1}S_{yx}S_{xx}^{-1}S_{xy})b = \rho^2 b \qquad [5]$$

The paired canonical variates (CVs), in matrices C_x and C_y, are conventionally ordered by decreasing correlation ρ. Given the theoretical Chi-square null distribution (e.g. the probability of zero correlation between variables of $\boldsymbol{X}$ and $\boldsymbol{Y}$), we iteratively estimate the net significance of components k to K via Bartlett's test (with Lawley's modification; Mardia *et al*., 1979), until the subset with non-significant correlation is identified.

For temporal CCA in fMRI data, voxels or principal component bases form the rows or features of $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be optimized, and C_x, C_y contain the time-series for which we are attempting to maximize correlation. As shown by Friman *et al*. (2002), this technique may be applied to estimate autocorrelation structure in a single dataset. For fMRI data matrix $\boldsymbol{X}$, we define time-shifted matrix $\boldsymbol{Y} = \boldsymbol{X}(t - \boldsymbol{\tau})$, for offset $\boldsymbol{\tau}$ (in units of TR, for fMRI). This produces weights a_k and b_k corresponding to the k^{th} set of voxels (or spatial bases) of highest autocorrelation. Taking the mean of paired a_k and b_k spatial weight vectors (for $1 \leq k \leq K$) produces a single CV matrix of maximally-autocorrelated time-series T_k. We term this technique "Canonical Autocorrelation Analysis" (CAA). The analysis is performed on a reduced PCA basis of the original $\boldsymbol{X}$ and $\boldsymbol{Y}$ matrices, to avoid rank-deficiency (since $n \gg t$), and to optimize the basis size K on which CAA is performed; note that we enforce the same subspace size K for both $\boldsymbol{X}$ and $\boldsymbol{Y}$. As a conservative upper limit, PC dimensionality should generally be restricted to $K < t/2$ (for t data timepoints), which ensures a non-degenerate solution to temporal CAA (see **5. APPENDIX** for the derivation).

2.3 Principles of CAA Denoising

The estimation of physiological noise is based on a set of known characteristics, as this noise has a specific spatiotemporal structure. In particular, we expect high temporal autocorrelation for spatially distributed signal (e.g. near regions of high vascularity), hence the use of CAA.

Physiological noise may also be characterized by power spectrum, as the effects of respiration and cardiac pulsation tend to be centered at frequencies of approximately 0.3 and 1.0 Hz, respectively, whereas BOLD oscillations are primarily localized in the 0.01-0.08 Hz spectral band (Zou et al., 2008; Cole et al., 2010). For standard fMRI acquisition TRs > 1 s, respiration effects are often partly aliased, though spectral power is generally concentrated above 0.1 Hz (Glover et al., 2000), while cardiac effects tend to be more extensively aliased. However, the undersampled cardiac effects have been shown to be higher-order processes (Penny et al., 2003) that comprise multiple oscillatory "modes", producing multiple spectral peaks for a given dataset (e.g. Kiviniemi et al., 2003). Data-driven analyses consistently detect high-frequency pulsatile components in data, with characteristic vascular-coupled spatial structure (e.g. McKeown et al., 1998; Kiviniemi et al., 2003; Perlbarg et al., 2006; Behzadi et al., 2007). Zollei et al. (2003) have also shown that for the majority of subjects, CAA extracts components with spectral peaks above 0.08 Hz, that have both vascular-coupled spatial patterns and correlations with external cardiac rate measurements. We therefore expect that spectral power > 0.10 Hz is predominantly due to physiological noise (as per Biswal et al., 1995; Zou et al., 2008; Cole et al., 2010). Although a portion of cardiac artifact often aliases below 0.10 Hz (Lund *et al*., 2006), and physiological noise also produces intrinsically low-frequency fluctuations (Birn et al., 2006; Shmueli et al., 2009), we conservatively focus on extracting high-frequency components for the current model, which is separable from BOLD signal based on spectral priors.

Finally, this structured noise originates from consistent brain regions, e.g. large vessels, sinuses and ventricles (Dagli *et al*., 1999; Windischberger *et al*., 2003; Birn *et al*., 2006), and thus regions of greatest physiological noise should be spatially reproducible. Based on these criteria, CAA is used to identify highly autocorrelated, high-frequency time series, in the PC-space dimensionality that maximizes the spatial reproducibility of these effects.

2.4 Algorithm Design

The PHYCAA denoising algorithm, shown schematically in Figure 1, consists of three segments: (1) *Component Selection*, from which we obtain PC timeseries on which to perform CAA, (2) *CAA optimization*, in which we estimate the most spatially reproducible dimensionality, and (3) *Physiological Noise Regression*, based on the estimated noise components. We begin by splitting a subject's fMRI

dataset temporally by run; for the present work, we apply the method to 2 runs, producing two 2D matrices of (*n* voxels x *t* timepoints), and a single reproducibility estimate. In practise, results may be extended to any number of runs $\geq$ 2, by maximizing the mean reproducibility for all pairwise comparisons of different runs, using the algorithm of Kettenring (1971), which was recently incorporated into the NPAIRS framework to increase reproducibility of activation maps (Afshinpour et al., 2010). Our algorithm is defined as follows:

Part 1 - feature selection: beginning with (*n* x *t*) data-split matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$, detrended using Legendre polynomial basis orders 0-3 (which optimizes average prediction and reproducibility for the datasets; see **2.7 Analysis and Model Performance** below for an overview of these metrics):

(1a) Remove task-design components: remove explicit task-design effects, as autocorrelated BOLD response may otherwise dominate global signal structure. Perform CVA on an optimized PCA subspace (i.e. Penalized Discriminant Analysis: PDA) of each split data matrix to maximally discriminate task-design labels, then perform an orthogonal projection from the task subspace, by regressing out the task CV scores from the PDA**,** to obtain residual matrices $\boldsymbol{R}^{(1)}_{\text{filt}}$ $\boldsymbol{R}^{(2)}_{\text{filt}}$. The optimal number of PC bases may be selected to jointly optimize prediction and reproducibility of the task subspace (Strother *et al*., 2010). For the presented data, we retained 40% (Trail-Making Test) and 20% (Sustained Attention to Response Task) of PC bases in each split matrix for signal estimation. These dimensionalities were empirically identified as providing the optimal mean performance metrics of prediction (and reproducibility, in the case of Trail-Making Test) across subjects, for a tested range of 10-50% (in increments of 10%); see **2.7 Analysis and Model Performance** for an overview of these performance metrics.

(1b) Singular Value Decomposition: apply SVD to matrices $\boldsymbol{R}^{(1)}_{\text{filt}}$ and $\boldsymbol{R}^{(2)}_{\text{filt}}$, to obtain orthogonalization $\boldsymbol{R}_{\text{filt}} = USV^{\text{T}}$ for each. This allows us to test of a range of dimensionalities, and thus optimize the spatial reproducibility of the estimated physiological effects.

Part 2 - CAA Optimization: perform iterative CAA on PC-space dimensionalities $3 \leq k \leq t/2$, to optimize the physiological noise timeseries for regression. We have generally observed, in the experimental data, that at least 3 PC dimensions are required to consistently capture a significant high-frequency component > 0.1 Hz. As previously noted, the upper dimensionality limit *t*/2 should be generally applied, to avoid a

potentially ill-conditioned solution (as specified in **5. APPENDIX**). For each iteration, CAA is performed on each split-half:

for k = [3 ... t/2], perform:

(2a) Principal Component Projection: project $\boldsymbol{R}^{(1)}_{filt}$ and $\boldsymbol{R}^{(2)}_{filt}$ into a basis of dimensionality k, by selecting the first k eigenvectors $U_k = [u_1 \ldots u_k]$ and computing:

$$\boldsymbol{Q}^{(1)}_k = U_k^T \boldsymbol{R}^{(1)}_{filt} \qquad \text{or equivalently,} \quad \boldsymbol{Q}^{(1)}_k = S_k V_k^T$$

And perform the same process to obtain $\boldsymbol{Q}^{(2)}_k$.

(2b) CAA decomposition: apply CAA independently on split-halves $\boldsymbol{Q}^{(1)}_k$ and $\boldsymbol{Q}^{(2)}_k$, via CCA of each matrix with its time-shifted version $\tau = 1$; we have tested time-shift range τ=1 to 5**,** and found no significant improvement in model performance for larger offsets. This step identifies the matrices $\boldsymbol{T}^{(1)}_k$ and $\boldsymbol{T}^{(2)}_k$ of most autocorrelated timeseries for each split.

(2c) Time-series Selection: identify a subset of relevant timeseries from each $\boldsymbol{T}_k$ matrix. The CAA components are ordered by decreasing autocorrelation (and associated significance). Retain only the $q < k$ components in which (1) all q timeseries are significant components and therefore highly autocorrelated, having $p < 0.05$ by modified Bartlett's test, and (2) the Discrete Fourier Transform (DFT) of the associated $\boldsymbol{T}_k$ timeseries have > 50% of the power spectrum above 0.10 Hz. From this selection step, we obtain reduced timeseries matrices $\boldsymbol{P}^{(1)}_{k,q}$ and $\boldsymbol{P}^{(2)}_{k,q}$.

(2d) Spatial Reproducibility: generate spatial maps of brain regions from $\boldsymbol{R}^{(1)}_{filt}$ and $\boldsymbol{R}^{(2)}_{filt}$ most associated with noise timeseries $\boldsymbol{P}^{(1)}_{k,q}$ and $\boldsymbol{P}^{(2)}_{k,q}$. This is obtained by computing the F-statistic of $\boldsymbol{P}_{k,q}$ fit at each voxel (e.g. the normalized ratio of explained variance from $\boldsymbol{P}_{k,q}$ / residual variance from $\boldsymbol{R}_{filt}$). This is not a "true" F-statistic, as such a measure assumes independence of data and regressors; since $\boldsymbol{P}_{k,q}$ is estimated globally from the data, individual voxels will have varying degrees of covariance with the $\boldsymbol{P}_{k,q}$ regressors. However, the pseudo-F provides a metric of the relative amount of estimated physiological noise variance at each voxel, generating weight maps $\boldsymbol{F}^{(1)}_{k,q}$ and $\boldsymbol{F}^{(2)}_{k,q}$. We then measure reproducibility $\boldsymbol{r}_{k,q}$ of the spatial noise distribution, via Pearson correlation between voxel values of $\boldsymbol{F}^{(1)}_{k,q}$ and $\boldsymbol{F}^{(2)}_{k,q}$.

Part 3 – Physiological Noise Regression: select the PC dimensionality k with maximum spatial reproducibility $\boldsymbol{r}_{k,q}$. Then remove noise-series estimates $\boldsymbol{P}^{(1)}_{k,q}$ and $\boldsymbol{P}^{(2)}_{k,q}$ from respective matrices $\boldsymbol{X}^{(1)}$ and

$\boldsymbol{X}^{(2)}$ via linear regression, to produce denoised data matrices $\boldsymbol{X}^{(1)}_{p}$ and $\boldsymbol{X}^{(2)}_{p}$ for further analysis.

In addition, we may use the pseudo-*F* noise-maps to produce a spatial map of regions most consistently influenced by physiological noise: taking $\boldsymbol{F}^{(1)}_{k,q}$ and $\boldsymbol{F}^{(2)}_{k,q}$, generate a scatterplot of pairwise voxel values. Then, project voxel values onto the primary axis of variance (1st PC dimension of the scatterplot), and normalize values by the standard deviation of the second axis of scatterplot variance (2nd PC dimension). From the projection, we obtain a normalized value corresponding to each voxel; this produces a Z-scored map of regions most consistently associated with physiological noise, analogous to the reproducible Z-scored statistical parametric maps (rSPM(Z)s) proposed by Strother *et al*. (2002; 2010).

2.5 Simulation Data

The PHYCAA method was applied to simulated data, to demonstrate artifact removal in a dataset with measurable "ground truth". We employed previously-established methods to simulate a block-design experiment with activation and baseline conditions (see Lukic et al. (2002) or Yourganov et al., (2011) for details). Images contained a 60x60 pixel "brain-like" image with additive Gaussian noise, consisting of "grey matter" in the center and rim of the phantom, and "white matter" in between. The amplitude of "grey matter" background signal was 4 times higher than in "white matter". Gaussian noise was spatially smoothed using a Gaussian filter with full-width-at-half-maximum (FWHM) of 2 pixels, producing noise standard deviation that was 5% of the background signal. Images during "activation" included 16 signal loci (12 in "grey matter", 4 in "white matter") with FWHM varied between 2-4 pixels, added to the smoothed noisy background image. We added 5 loci of 2-4 pixel FWHM to simulate large vessel regions, in which pulsatile bloodflow effects tend to occur. In addition, we obtained a spatial gradient map of the temporal variance image; these edge regions were used to simulate the effects of motion and susceptibility changes due to respiration (which occur near brain edges). Fig. 7(a) shows signal and artifact loci in the phantom. A mask with J=2072 pixels was used to exclude locations outside of the phantom from analysis.

The images were arranged into 10 epochs of 20 images to simulate a block design; each epoch included 10 "activation" images followed by 10 "baseline" images (200 scans total, per sample). To simulate the hemodynamic response, each pixel's time course was convolved with a hemodynamic response function (HRF) defined by the sum of two Gamma functions, using parameters obtained from

Worsley (2001). Amplitudes of the activation loci were sampled from a multivariate Gaussian distribution. The mean activation amplitudes scaled to 1.5% of local background signal; with HRF convolution, this corresponds to contrast-to-noise ratio (CNR) 0.5. The variance of the activation amplitude was set equal to the variance of background Gaussian noise added to each voxel, including activation voxels, producing greater noise variance in activation regions. We also defined the covariance between Gaussian activation signal amplitudes at locations *j* and *k* by $(\sigma_j\sigma_k)\rho$. We set $\rho = 0.5$ to approximate typical functional connectivity for a distributed spatial network (Lukic et al., 2002).

We simulated physiological artifact in the datasets using Trails-Making Test experimental data (see ***2.6.1 Trail-Making Test (TMT)*** below), as this block design is similar in temporal structure to the simulation model. We first determined the expected ratios of cardiac/respiratory variance, relative to background variance: for each subject, we selected the 5% of voxels showing greatest change in variance when denoised using only cardiac regressors in RETROICOR, as loci of cardiac noise. We then measured mean variance σ_{card}^2 in these voxels, and mean variance σ_0^2 for the remaining voxels, designated "background voxels". Assuming approximate independence of the cardiac and background signals (e.g. additive variances), we then estimate the ratio of cardiac/background variance by:

$$R_{card} = (\sigma_{card}^2 - \sigma_0^2) / \sigma_0^2 \qquad [6]$$

This process was repeated for respiratory regression, to obtain ratio R_{resp}. We obtained mean R_{card} of 3.54 (range 2.22 to 4.46) and mean R_{resp} of 3.88 (range 1.97 to 5.13). To simulate respiration timecourses, we used the raw output from the pneumatic belt, convolved with a 1s smoothing filter to remove local spikes. To simulate pulsatile flow effects, we interpolated a cosine function between successive cardiac peaks. To generate physiological artifact in a given dataset, we then randomly selected cardiac and respiratory waveforms from the 19 subjects in the experimental dataset, and embedded them into their respective artifact loci, with variance rescaled to the mean R_{card} or R_{resp}, at each voxel. For our analyses, we generated 100 such simulation datasets.

2.6 Experimental Data

We performed analyses based on two datasets from a cognitive task battery, designed for eventual clinical implementation in the assessment of stroke and vascular-cognitive impairment. Nineteen young, healthy volunteers participated in the study (11 female, ages 21-34 yrs, mean 26.5 yrs). Subjects were confirmed right-handed using the Edinburgh Handedness Inventory (Oldfield *et al*., 1971), and screened for cognitive and neurological deficits, by self-report and using the Mini-Mental Status Examination (Folstein *et al.*, 1975), with group mean 29.5 and scores ranging from 28 to 30 (out of 30).

The BOLD fMRI data were acquired on a 3T MR scanner (MAGNETOM Tim Trio, VB15A software; Siemens AG, Erlangen, Germany), with a 12-channel head coil. A T1-contrast anatomical scan was obtained (oblique-axial 3D MPRAGE, 2.63/2000/1100 ms TE/TR/TI, 9° FA, 256 X 192 matrix, 160 slices per volume, voxel dimensions 1x1x1 mm^3), followed by BOLD fMRI (2D gradient-echo EPI, 30/2000 ms TE/TR, 70^o FA, 64x64 matrix, 30 slices per volume, voxel dimensions 3.125x3.125x5 mm^3).

During scanning, we also measured cardiac and breathing rates, via photoplethysmograph and pneumatic belt, respectively. Subjects received a 15 minute orientation session in an MRI simulator, and performed two runs of each task in the scanner, each separated by approximately 10 minutes of other behavioural tests. The tasks included (1) a block-design adaptation of the behavioural Trail-Making Test (TMT), and (2) a rapid event-related Sustained Attention to Response Task (SART). These tasks allow the testing of interactions between task design with physiological noise; the TMT design is characterized by multiple levels of cognitive engagement and subject-paced tracing tasks, whereas SART is driven by rapid sequential stimuli, requiring consistent, rapid response (and thus attention) to maintain performance. As both tasks use data from a fixed set of subjects, acquired during the same sessions, over comparable scanning times, we are able to directly examine the influence of these task design factors on the identified physiological noise structure.

2.6.1 Trail-Making Test (TMT)

The task consists of 2 types of stimuli: *TaskA*, in which numbers 1-14 are pseudo-randomly displayed on a viewing screen, *TaskB*, in which numbers 1-7 and letters A-G are displayed. Subjects drew a line connecting items in sequence (1-2-3-4-…) or (1-A-2-B-…), respectively, connecting as many as possible for a 20s block interval, while maintaining accuracy. A *Control* stimulus was presented after each task block, in which subjects trace a line from the center of the screen to a dot (randomly placed at a

fixed radius from the center of the screen) and back over 2s, repeated 10 times. Subjects performed a 4-block, 40-scan epoch of *TaskA–Control-TaskB-Control* 2 times per run, with 2 runs per subject. Tracing was performed with an MRI-compatible writing tablet and stylus, with performance monitored on a projection screen (Tam *et al*., 2011). The timeseries of runs 1 and 2 formed matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$ for PHYCAA; given the t=80 timepoints of each run, t/2=40 provides the theoretical upper limit on CAA dimensionality for this design. For this dataset, we analyzed the contrast between Task B and Task A conditions without the control state, as this weaker cognitive contrast is likely to be more sensitive to physiological noise.

2.6.2 Sustained Attention to Response Task (SART)

This task was presented as a fast event-related GO-NOGO design (Fassbender *et al*., 2004). The set of integers 1-9 were presented in random order on the screen, followed by a masking image. Stimuli were presented for 250 ms, while the mask was shown for a randomized interstimulus interval, of mean 1250 ± 210 ms. Subjects were asked to respond to all integers except '3' (the NOGO stimulus) using the MRI-compatible writing tablet, by touching the stylus to the tablet surface. A single run consisted of 100 presented numbers, with 75 GO stimuli and 25 NOGO stimuli, in randomized order, with 76 scans per run. We acquired 2 runs of this task per subject. As with TMT, the full run of t=76 scans was used for PHYCAA denoising, giving the theoretical dimensionality bound t/2=38 for CAA analyses. For task analysis, we compared correct GO response (touching the tablet surface) to correct NOGO response (motor inhibition). In order to minimize bias due to excess GO conditions, we randomly sampled 25 GO scans to balance the 25 NOGO scans for each run, when performing the analysis.

2.6.3 Data Preprocessing

Preprocessing methods were applied prior to CAA denoising, primarily AFNI utilities (Cox, 1996); this was based on the optimal fixed preprocessing for the task design, as shown by Churchill *et al*. (in press). Rigid-body motion correction was performed with *3dvolreg*, using a weighted least-squares cost function and Fourier interpolation; all volumes were registered to the 10^{th} scan-volume of run 2. We then performed slice-timing correction with *3dTshift* via Fourier interpolation, followed by in-plane spatial smoothing using *3dmerge* and a 6.0 mm FWHM Gaussian kernel. Detrending was performed using *3dDetrend*, for Legendre polynomial orders 0-3. The EPI and T1 brain masks were obtained for each

subject, using the FSL Brain Extraction Tool (Smith *et al*., 2004). In order to perform intersubject comparisons, 4D EPI data were transformed into a common space prior to analysis, using a group-defined template obtained with the FSL *flirt* registration utility. Spatial normalization of EPI data were obtained by aligning subject T1s using an iterative process: (1) register T1 volumes to MNI 152 (Montreal Neurological Institute template); (2) average the masked, registered T1 brain volumes; and (3) re-register T1 volumes to the normalized average. Steps (2-3) were repeated three times to optimize inter-subject alignment (Guimond *et al*., 2000).

2.7 Analysis and Model Performance

To demonstrate the generalizability of results, we also tested CAA denoising on two different analysis models. For TMT and simulation data, we used a multivariate Penalized Discriminant Analysis (PDA) model built on an optimized Principal Component (PC) basis; for SART, we performed searchlight Gaussian Naïve Bayes (slGNB), a locally-multivariate predictive analysis method, as the basic GNB model generally produced poor discrimination for this task. The methods are briefly summarized; in both cases, we performed 2-class analyses, in which each class consists of scans acquired under different stimulus conditions (*Task B* vs. *Task A*, or *GO* vs. *NOGO*).

Given that fMRI data has no direct measure of "ground truth" (e.g. whether activations are real or artifactual), we require an alternate means of quantifying model performance. Strother et al. (2002; 2010) have proposed the data-driven metrics of temporal prediction accuracy (*P*) of brain-state, and the reproducibility (*R*) of SPMs, obtained in a split-half cross-validation framework, as effective metrics of preprocessing effect on the underlying true data. These measures quantify important characteristics of the fMRI data, namely, whether results generalize to independent datasets (via prediction across splits), and the consistency of identified brain region signals (via reproducibility). They also provide a target for model validation, as an effective preprocessing model should, at a minimum improve the predictive accuracy and/or reproducibility of fMRI data.

For the block-design analyses, we performed 2-class PDA in the NPAIRS split-half resampling framework (Strother *et al*. 2002, 2010) to obtain (P,R) performance metrics, along with a reproducible, Z-scored SPM (rSPM(Z)); the process is described as follows. PCA decomposition was first applied to the (*n*-voxel x *t* timepoint) data matrix, for dimensionality reduction and denoising; the dataset was then split

into halves, and we performed PDA on each split-half. To do so, we performed a second PCA, and selected the common PC basis size that jointly optimized prediction and reproducibility of results. In the reduced PC-space, the vector was computed which, for within-class covariance matrix $\boldsymbol{W}$ and between-class covariance matrix $\boldsymbol{B}$, maximized $\boldsymbol{W}^{-1}\boldsymbol{B}$ (e.g. the eigenvector of $\boldsymbol{W}^{-1}\boldsymbol{B}$). This eigenvector is projected into voxel-space, to produce a normalized statistical parametric map (SPM), for each split-half, that discriminates between experimental conditions.

The *R* value was computed by Pearson correlation between the two SPMs. An rSPM(Z) was obtained by projecting the Z-scored SPMs onto the axis of maximum correlation (signal axis), and normalizing by the orthogonal axis (noise axis). For *P*, we used the model parameters of a training dataset (split-half 1) to predict the experimental condition of scans in a test set (split-half 2) and vice-versa, using posterior Bayes probabilities calculated per scan. In TMT, the 4 *TaskA*-*TaskB* blocks obtained from both runs allows for 3 independent split-half resamples for each subject (given 3 unique ways of separating the 4 blocks into two groups of 2), afterwards computing mean (*R*,*P*) across splits, and the mean rSPM(Z). For consistency, we also applied this resampling scheme to simulation data.

We also performed slGNB for 2-class estimation of GO vs. NOGO conditions. To analyze results with the slGNB model (Kjems *et al*., 2002; Pereira *et al*., 2009), we defined a spherical cluster of radius 6.25 mm (e.g. 2 voxels in the axial plane), at each voxel (Kriegeskorte *et al*., 2006). For each voxel cluster, we then trained a Naive-Bayes classifier on the training dataset, and used the classifier to label samples from the test set. The run-1 data were used to classify scans in run-2 and vice-versa, obtaining 2 estimates of prediction accuracy at each voxel; we then computed mean voxelwise prediction accuracy, generating a spatial brain map of prediction. Classification was performed on the scan at the optimal delay after each GO/NOGO stimulus, for 50 scans per run (25 GO / 25 NOGO scans). We optimized the event-related design by identifying the lag, of TR = 3 to 5 in steps of 1, that maximized the 95th percentile of voxel prediction values specific to each subject. This allows us to estimate the lag at which peak haemodynamic response occurs.

2.8 Characterizing Autocorrelation Structure in fMRI Data

We first examine model stability as a function of PC dimensionality, to determine whether experimental results agree with the theoretical predictions. Model stability is dependent on the condition

number of covariance and cross-covariance matrices, that is, their sensitivity to minor perturbations, as they are used to compute the CAA transform. We plot the L_2-norm condition number, the ratio of largest/smallest eigenvalue, of matrix S_{xy} (used in Eqtn.s [5,6]) as a function of PC dimensionality.

We then examined the spectral structure of CAA timeseries components, for each task. An intensity map was plotted, in which pixel intensity represents the fraction of power-spectrum > 0.10 Hz (e.g. the physiological noise band) for each timeseries CV (as estimated in algorithm step 2(c)), at a given PC dimensionality. The CVs are ordered by decreasing autocorrelation of τ to τ+1; we display the mean of intensity values, computed over all subjects. In addition, the number of CVs with $p \leq 0.05$ threshold is plotted as a function of PC dimensionality. This demonstrates the inter-subject variability in the number of significant noise CVs, at each PC dimensionality.

2.9 Physiological Noise Reproducibility and PC-space Optimization

Trends in PC dimension were also examined as a function of pseudo-*F* map reproducibility and global signal-to-noise (gSNR). For each subject, we determine the PC dimension *k* that maximizes spatial reproducibility r_k of the estimated physiological noise sources. We may compute the spatial gSNR of the estimated physiological-noise signal in the brain, as gSNR is monotonically related to reproducibility r_k. The formula defined by Yourganov *et al*. (2011), of $gSNR = \sqrt{2r/(1-r)}$, expresses gSNR as a function of correlation r_k. We plotted subject gSNR relative to PC dimension, for optimized subject reproducibility.

In addition, we tested for a relationship between PHYCAA-derived physiological noise and trends in externally-measured cardiac and respiratory cycles, to determine how these processes potentially influence the estimated physiological noise. It is hypothesized that increased variability in cardiac and respiratory processes, and thus more temporally complex, aperiodic physiological effects, is associated with an increase in the required PC dimensionality to obtain an optimal estimate of physiological noise in PHYCAA. Cardiac rate values were computed as 1/(time interval) between all successive pulse pairs; as a conservative correction for sampling errors, we rejected samples (e.g. single interval measurements) that are both more than 3 standard deviations from the mean, and outside a conservative range of 0.83-2.33 Hz (from 40-140 bpm, or low-range resting rate, up to 70-80% of HR_{max}, e.g. Taylor *et al*., 1969).

For respiration, the rate and depth potentially account for separate BOLD signal effects (Windischberger *et al*., 2003; Birn *et al*., 2006), and are thus examined. We convolved the respiratory

waveform with a 0.5 s averaging filter to control local fluctuations, and identified peaks in successive inspiration/expirations, defined as extrema in the deflections above/below the mean respiratory amplitude. The respiratory rate was computed via 1/(time interval) between successive inspiration peaks. For respiration amplitude, we converted the respiratory waveform to percentage of the full range (e.g. maximum – minimum displacement), and measured the % displacement from successive expiration/inspiration peak pairs. We computed the mean of respiratory and cardiac rates, and standard deviation (SD) for both rates and respiratory amplitude; mean respiratory amplitude is omitted, as it is a relative measure. We tested each of these measures for significant association with optimized subject PC dimensionality using the nonparametric Spearman correlation. The 95% confidence interval (CI) was obtained for the correlations, using Bootstrap resampling (1000 iterations).

Additionally, we plotted the brain regions that are most consistently associated with CAA-estimated physiological noise, by computing the voxelwise mean of the Z-scored physiological noise maps, over all subjects; the maps are shown for individually optimized PC bases. This demonstrates both the spatial consistency and mean signal strength of the estimated physiological noise in brain regions.

2.10 Effects on Analysis Results

We then examined the effects of PHYCAA on analysis results, and compared to the effects of the commonly-used RETROICOR method of physiological noise regression (Glover *et al*., 2000). This method uses measurements of cardiac and respiratory rate to regress out Fourier-series approximations of physiological noise. For simulation data, we counted true positive / false positive rates (TPR/FPR) in "gray matter" of signal+artifact data, for a range of Z-score thresholds. We computed TPR, estimated across all 100 "subjects", for each of the 12 peak gray-matter signal loci; FPR was similarly computed in 12 randomly selected background pixels, which are more than FWHM (of the smoothing filter) away from any signal or artifact regions. We then plotted the mean Receiver Operating Characteristic (ROC) curves. This allowed us to compare *relative* signal detection changes from performing PHYCAA and RETROICOR, in the simulated mixture of signal and noise. We plotted ROC curves for (1) Gaussian noise only (as a baseline reference), (2) with cardiac and respiratory artifact, and cardiac/respiratory artifact with (3) PHYCAA denoising, and (4) RETROICOR denoising. We also plotted false positive rates

in the peak cardiac and respiratory artifact loci, as a function of Z-score, to compare denoising performance in these regions.

For TMT data, we measure the changes in global (*P, R*) metrics due to PHYCAA denoising compared with an optimal fixed pipeline without physiological denoising, as well as RETROICOR denoising. We also examine the effects of PHYCAA on spatial activation maps. For reference, we compare the mean Z-scored SPMs before/after PHYCAA denoising. Regions of significant change are identified via Bootstrap estimation of the mean change (1000 iterations). This provided the probability of significant positive/negative change at each voxel, corrected for multiple comparisons via False-Discovery Rate (FDR) 0.10; this threshold is applied as a more inclusive estimate of significant change, due to the relatively high between-subject heterogeneity for these single-subject results (see **3.3 Effects on Analysis Results** below). In addition, we applied a conservative cluster-size threshold of 3 voxels to the regions of significant change, which improved the interpretability of the difference maps.

For SART, we determined the effect of PHYCAA on voxel prediction, by computing the percent change in suprathreshold voxels, for a range of % accuracy thresholds. The percent change in suprathreshold voxels is also plotted for RETROICOR relative to minimal preprocessing, to demonstrate differences in prediction between the two models. In addition, we depict the effects of PHYCAA on spatial prediction maps. For reference, we display mean prediction accuracy maps, computed over all subjects, for basic preprocessing (no physiological noise correction) and PHYCAA denoising. We again estimate probability of significant change in prediction accuracy at each voxel, via bootstrap estimates at each voxel (1000 iterations). The mean prediction change is then plotted for regions showing significant positive/negative difference at FDR=0.10 (chosen due to the relatively high inter-subject heterogeneity of SPMs), again applying a cluster threshold of 3 voxels to improve interpretability.

As an additional test, we examined whether the removal of task-design components prior to estimating physiological noise (**2.4 Algorithm Design,** Step 1(a)) was the cause of PHYCAA's improved performance over RETROICOR, since this protects against the removal of task-related BOLD spectral power. A modified version of RETROICOR was performed, in which we first regressed out the task-component using CVA (as in Step 1(a) of the PHYCAA algorithm), then used the residual data to estimate physiological noise regressors, which were subtracted from the initial data matrix prior to analysis. We

then compared the performance of the modified RETROICOR procedure against PHYCAA, for both TMT and SART tasks. In addition, we examined whether cardiac and respiratory phases tended to correlate with the task response, to validate any observed effects of modified RETROICOR. For each subject and task, we measured correlation between the estimated task response and undersampled cardiac/respiratory regressors, for a range of time-shifts (in units of TE ≈ 0.06, up to 1 TR); we retained the time-shift giving maximum possible correlation between cardiac/respiratory phase and task design, for each subject, in both TMT and SART tasks.

3. RESULTS

3.1 Characterizing Autocorrelation Structure in fMRI Data

Fig. 2 plots the L_2 matrix condition number for cross-covariance S_{xy} as a function of PC dimensionality, for the TMT experiment. Condition tends to increase monotonically with PC dimension, until a transition of increased condition at PC > t/2 (40 PCs). The condition number distribution then remains relatively consistent until a second transition at PC = 70, at which point the S_{xy} matrices become rapidly more ill-conditioned. CAA thus becomes more unstable beyond the PC > t/2 cutoff, although it is not until the PC = t limit is approached that condition becomes completely unstable, with an asymptotic increase in matrix condition number.

The autocorrelation structure of the CAA components is displayed graphically in Fig. 3, for both tasks. Fig. 3(a,b) show the fraction of the temporal power spectrum > 0.10 Hz across subjects (e.g. the physiological noise frequency band) for each CV (vertical axis), as a function of initial PC dimensionality (horizontal axis). Each pixel represents mean fraction of the power spectrum above 0.10 Hz, averaged across all subjects, for a given CV. The CVs are ordered by decreasing autocorrelation, and the black curve on the plot marks the median p=0.05 autocorrelation significance cutoff at each PC dimensionality. Fig. 3(c,d) boxplots show the number of significant CVs (p=0.05) as a function of PC dimension, across all subjects. The dashed line represents the identity line, at which (# significant CVs = PC dimensionality). For all plots, the t/2 dimensionality limit is plotted as a vertical black line.

The spectral plots, displayed in Figs. 3(a,b), show similar structure for both tasks. Below the black line, CVs are not significant (p < 0.05) and have a relatively high fraction of spectral power > 0.10 Hz.

Above the black line CVs are significantly autocorrelated ($p < 0.05$), and have a generally lower fraction of spectral power > 0.1 Hz. For PC dimensions < t/2 (40 PCs for TMT, 38PCs for SART), there are a subset of identifiable CVs that attain both greater than 50% of spectral power above 0.10 Hz, and significant autocorrelation (i.e. blue-green to yellow pixels, in a band above the black line). The number of significant CVs with more than 50% of power above 0.10 Hz tends to be relatively consistent across subjects, for a given PC dimensionality (Fig. 3(c,d)). However, for PC dimensions > t/2, a transition gradually occurs, as mean spectral power is now observed to be more uniform across CVs, with increasing PC dimensionality. For these dimensionalities, an increasing proportion of components account for significant, high-frequency autocorrelation, rapidly increasing the possible number of identified physiological regressors (Fig. 3(c,d)).

3.2 Physiological Noise Reproducibility and PC-space Optimization

Figure 4 plots subject PC dimensionality against gSNR of the estimated physiological *F*-maps with optimized reproducibility. For all subjects, the optimum is consistently below the PC dimensionality t/2 transition. Comparing task-designs, TMT has a wider range of optimal PC bases than SART, generally requiring higher dimensionality to optimize, with respective means 11.8 ± 8.5 and 6.2 ± 3.5. From these PC bases, we also identify an average of 4.9 ± 4.0 significant CVs, or timeseries components, for TMT, and 2.4 ± 2.1 for SART. TMT also consistently produces much lower gSNR than SART, with respective means of 1.25 ± 0.50 and 11.44 ± 13.56. This is associated with lower spatial reproducibility of physiological noise, as TMT has mean 0.417 ± 0.181, whereas SART has mean 0.922 ± 0.079. We thus observe task-specific curves, in which decreased physiological gSNR corresponds to an asymptotic increase in PC dimensionality.

We also examined the relationship between the dimensionality of CAA-estimated physiological noise and externally-measured cardiac and respiratory processes, using Spearman correlations; the results are summarized in Table II. For both tasks, dimensionality of the noise subspace is non-significantly correlated with mean cardiac/respiratory rates. By contrast, dimensionality is significantly correlated with SD of respiratory rate (Figure 5(a,d)) and amplitude (Figure 5(b,e)). The similarity of trends for SD of respiratory rate and amplitude is not unexpected, as these processes are themselves significantly correlated, at 0.76 (p<0.001) and 0.54 (p=0.022) for TMT and SART, respectively. The PC

dimensionality produces weaker correlations with cardiac SD, of 0.32 (95% CI: -0.21 to 0.78) and 0.53 (95% CI: 0.08 to 0.63) for TMT and SART. However, in each plot, 2 subjects with highest PC dimensionality (white circles) reinforce a positive correlation with respiratory SD, but negatively impact cardiac SD correlations. These points were tested using the Cook metric D_{cook} (Cook & Weisber, 1982), which measures potential outlier influence on the polynomial fit, and were found to be significant outliers, based on the recommended $D_{cook}>4/n$ threshold (Bollen & Jackman, 1990). If we remove the outliers, correlation between PC dimensionality and cardiac SD consistently increases for both tasks (see Table II); of the 2 identified outliers, one represents the same subject across both TMT and SART.

Comparing between tasks, TMT tends to produce higher mean respiratory rate than SART, with respective means 0.399 ± 0.051 Hz and 0.364 ± 0.043 Hz (p<0.001 by paired Wilcoxon test), and higher respiratory rate SD, with respective means 0.069 ± 0.019 Hz and 0.044 ± 0.014 Hz (p<0.001, paired Wilcoxon). TMT produces higher SD in respiratory amplitude than SART, with means (12.84 ± 2.68)% and (10.20 ± 2.18)% (p<0.001, paired Wilcoxon). The TMT also generates higher cardiac rates than SART, with means 1.219 ± 0.163 Hz and 1.175 ± 0.160 Hz respectively (p<0.001, paired Wilcoxon), along with higher cardiac SD, with means 0.096 ± 0.039 Hz and 0.085 ± 0.042 Hz (p=0.106, paired Wilcoxon). Therefore, the higher PC-dimensionality TMT data again evidences greater associated physiological variability.

Figure 6 displays spatial regions consistently identified by the PHCAA technique, via Z-scored physiological noise maps, averaged over all subjects. For both designs, CAA predominantly identifies signal in the brainstem (slice 9), major arterial tracts including the Circle of Willis (slices 21-27) and middle cerebral arteries (slice 35), as well as the straight sinus (slice 35), frontal sinus (slice 60-69) and lateral ventricles (slices 35-51). In addition, PHYCAA identifies artifactual signal in gray matter regions, more visible for SART in Fig. 6, such as the superior temporal lobes (slice 35) right middle frontal (slice 51) and precuneus (slice 69). However, the highest Z-scores for physiological noise in SART tend to be consistently larger than for TMT. The average 95th percentile of subjects' noise map Z-scores is 5.79 ± 1.98 and 24.2 ± 22.0 for TMT and SART, respectively; all subjects show increases in both the 95th percentile, and overall Z-score distribution (p<0.001, nonparametric Kolmogorov-Smirnov test for equality

of cumulative distributions). This is consistent with generally higher gSNR of the estimated noise spatial structure in SART.

Furthermore, the task-dependent difference in physiological noise Z-scores is primarily due to improved reproducibility of the pseudo-*F* maps, and not higher absolute pseudo-*F* statistics. Comparing subject pseudo-F maps between the two tasks, 12/19 subjects show consistently higher pseudo-*F* statistics for TMT than SART, whereas for 7/19 subjects the converse is true (all significant at $p<0.001$, Kolmogorov-Smirnov test). In addition, the average 95^{th} percentile of F-statistics is 16.03 ± 16.47 and 22.65 ± 19.91 for TMT and SART, respectively ($p=0.171$, Wilcoxon paired difference test). There is thus no significant difference between SART and TMT tasks in the relative fraction of variance attributed to physiological noise, whereas the relative variance shows consistently higher spatial reproducibility for SART.

3.3 Effects on Analysis Results

Figure 7 depicts the effects of PHYCAA on signal detection in simulated data. Fig. 7(a) shows the spatial distribution of signal and artifact loci. Fig. 7(b) plots the ROC curves for (1) Gaussian noise only, (2) simulated physiological artifact, and with both (3) PHYCAA and (4) RETROICOR denoising; 95% CI bars are also shown. The addition of physiological artifact significantly reduces signal detection; signal detection is not significantly better than the line of identity TPR=FPR, for higher thresholds (FPR<0.05). The use of PHYCAA significantly improves signal detection, as the 95% CI generally do not overlap with uncorrected data, although mean signal detection is consistently lower than for Gaussian-only data. Although RETROICOR consistently improves mean signal detection, the 95% CIs overlap with uncorrected data. Fig.s 7(c-d) show FPRs in cardiac and respiratory artifact loci, again with 95% CI bars. For both artifacts, PHYCAA significantly reduces artifact Z-scores, while RETROICOR has a significant, but weaker impact on physiological artifact.

In Figure 8, we show the effects on prediction (*P*) and reproducibility (*R*) when applying PHYCAA; we also compare these effects with RETROICOR preprocessing. For all subjects, PHYCAA denoising improves either *R* or *P*. For PHYCAA, mean *R* change 0.091 ± 0.090 (range -0.031 to 0.250) is observed with 16/19 subjects improved. Subjects also demonstrate mean change in *P* 0.065 ± 0.062 (range -0.037 to 0.157), with 16/19 subjects showing improvement. By comparison, RETROICOR has inconsistent

effects on model performance: mean *R* change is -0.047 ± 0.133 (range -0.357 to 0.201), with 6/19 subjects improved, and mean change in *P* of -0.012 ± 0.071 (range -0.184 to 0.079), with 9/19 subjects improved.

In Figure 8(c) we plot slices depicting the net effects of PHYCAA denoising on SPMs; mean Z-scores across individually analyzed subject SPMs are shown with/without PHYCAA applied, along with mean signal change, significant at FDR=0.10 under bootstrap estimation. The subject SPMs are relatively heterogeneous, with intersubject SPM correlations ranging from -0.12 to 0.36 (basic preprocessing, of no physiological noise correction) and -0.08 to 0.29 (PHYCAA denoising)

Shown in Fig. 8(c), top/middle, the strongest task-positive activations (elevated signal in Task B) are observed in the right cerebellum (slice 12), left middle/superior temporal and occipital lobes (slices 38, 44), left inferior frontal lobes (slice 44), left precentral gyrus, superior parietal lobes and precuneus (slice 63), the superior parietal lobes, left middle frontal lobe and supplementary motor area (SMA) (slices 63, 72). Task-negative activation (elevated signal in Task A) is observed in the ventral anterior cingulate cortex (vACC) and right middle temporal lobe (slice 38), posterior cingulate cortex (PCC) and superior medial-frontal lobe (slice 44), and right supramarginal and postcentral gyri (slices 63, 69).

Regarding the PHYCAA denoising effect (Fig. 8(c), bottom), PHYCAA denoising reduces mean signal in the medial ventricles (slice 38), as expected. In addition, it consistently increases Z-scores in a subset of high-signal regions including the right cerebellum (slice 12), left superior temporal lobe (slices 38, 44), the left inferior frontal lobe (slice 44), left superior parietal lobe, SMA and precentral gyri (slices 63), as well as task-negative PCC (slice 44) and right postcentral gyrus (slices 63, 69). PHYCAA denoising also appears to weakly but consistently increase signal in right-side medial regions although they retain sub-threshold mean Z-scores (e.g. in slices 38,44 of denoising effect images). This demonstrates that PHYCAA tends to increase mean signal strength in a set of brain regions with highest common Z-scores.

Figure 9 demonstrates the effects of PHYCAA and RETROICOR denoising on analysis of SART. In Figure 9(a), we plot the net change in the number of voxels above a given prediction accuracy threshold, for PHYCAA and RETROICOR, relative to basic preprocessing (no physiological noise correction), for all subjects. PHYCAA results generally show an increase in predictive voxels. For the

shown threshold range, only two subjects consistently show a weak decrease of <2% in predictive voxels; all other subjects show consistent improvement, defined as an increase in the number of voxels that are predictive of brain-state for a given threshold, with a consistently positive median change in prediction accuracy. By comparison, RETROICOR has a wider range of prediction effect (positive and negative) across thresholds, however the effect on model performance is more frequently detrimental, with 10/19 subjects' prediction consistently reduced for all thresholds, and a generally negative median change.

Figure 9(b) depicts the effects of denoising on spatial prediction maps; subject SPMs are highly heterogeneous for this task as well, with intersubject SPM correlations ranging from -0.13 to 0.14 (basic preprocessing) and -0.12 to 0.23 (PHYCAA denoising). The Fig. 9(c) top/middle plots show mean prediction maps with/without PHYCAA; this demonstrates regions that are predictive of the GO vs. NOGO task contrast. The most predictive regions are located in the left inferior/middle frontal and right superior frontal lobes (slice 45) and right insula (slice 39), with weaker predictive regions including right middle-frontal, left superior temporal (slice 39), left middle occipital and inferior frontal (slice 45), right-middle and superior medial frontal and inferior parietal lobes (slice 58,65), right precentral (slice 65), and superior frontal, SMA and precuneus regions (slice 72). We also display brain regions showing consistent prediction change, with probability given by Bootstrap estimation, corrected for multiple comparisons at FDR=0.10. Based on these criteria, 7.26% of brain voxels show significant prediction increase, while 0.12 % of voxels become consistently less predictive. The mean prediction change due to PHYCAA is shown in Figure 9(b) bottom; with the 3-voxel clustering threshold, 1.93% of voxels show significant prediction increase, and 0 voxels show consistent predictive decrease. Significant predictive increases of greatest spatial extent include left superior temporal (slice 39), left inferior frontal (slice 45), superior medial frontal and right inferior parietal (slices 58, 65), superior frontal, SMA and precuneus regions (slice 72).

As a final test, we performed modified RETROICOR, in which the physiological noise regressors were estimated from residual data. For TMT, the modified RETROICOR shows no consistent improvement relative to standard RETROICOR, with mean R change 0.023 ± 0.105 and mean P change 0.013 ± 0.077 (p=0.617 and 0.528 respectively, paired Wilcoxon). The change in (P,R) relative to minimal preprocessing also remains suboptimal to PHYCAA: mean R change is -0.025 ± 0.129 (range -0.264 to 0.195), with 7/19 subjects improved, and mean change in P of -0.001 ± 0.094 (range -0.165 to 0.146),

with 10/19 subjects improved. Similarly for SART, regressing out the estimated task response has no significant impact on the effects of RETROICOR, with 10/19 subjects again showing prediction decreases and a consistently negative median prediction change, compared to standard preprocessing. It is not surprising that removal of the task component has relatively little effect on RETROICOR's performance, as the downsampled cardiac and respiratory phases are generally uncorrelated with task design. For TMT, cardiac and respiratory regressors evidence maximum correlation with task design of 0.11 ± 0.06 and 0.10 ± 0.09, respectively (mean ± S.D. across subjects); for SART, cardiac and respiratory regressors evidence maximum correlation with task design of 0.12 ± 0.13 and 0.06 ± 0.03, respectively.

4. DISCUSSION

The presented algorithm performs multivariate, data-driven characterization and removal of physiological noise in fMRI data. As we have demonstrated, the technique provides an effective method of artifact removal, without requiring external measurements, such as cardiac and respiratory rates, or classification based on spatial priors (e.g. Perlbarg *et al*., 2007; Tohka *et al*., 2008). In addition, this method outperforms the parametric physiological modelling of RETROICOR, in both experimental and synthetic datasets. Our results demonstrate that PHYCAA's advantage does not depend on regressing out the task effect prior to performing CAA, as RETROICOR gains no significant advantage with the same orthogonalization. This is consistent with our measured cardiac/respiratory phase correlations; although the rate and variance of cardiac and respiratory processes correlate with cognitive engagement (Foster & Harrison, 2004; Birn et al., 2010), we show that the phase values are generally uncorrelated, and thus RETROICOR is insensitive to task-coupling in these datasets. It is unclear why RETROICOR remains detrimental to a set of subjects, although it is possible that the phase-matching process is fitting to, and removing, transiently task-linked fluctuations that are important for multivariate classification.

The PHYCAA method also avoids an additional issue of parametric techniques, highlighted by Jones *et al*. (2008) and Churchill *et al*. (2010), of a dependency on the ordering relative to other preprocessing steps, as we estimate noise regressors directly from fMRI data. We thus have an estimation method that does not require hand-selection of components as often used in ICA, and reduces issues of signal/noise mixing, by explicit decorrelation with the task design.

We examined the physiological-signal subspace identified by PHYCAA, which evidenced a relationship between spatial reproducibility and dimensionality, analogous to that of neuronally-coupled BOLD signal (Fig. 4). Decreased gSNR of the identified sources is associated with an asymptotic increase in estimated dimensionality; this mirrors the results of Yourganov *et al*. (2011), who found similar trends in multivariate BOLD fMRI analyses, where increased dimensionality (and lower reproducibility) is associated with decreased grey-matter functional connectivity.

Similarly, this high dimensionality / low reproducibility trend is shown to be associated with increased SD in respiration and cardiac rates, and respiratory amplitude variability (Fig. 5), but generally uncorrelated with mean respiratory and cardiac rates. This demonstrates that the PHYCAA method is at least partially capturing the heterogeneous, subject-dependent effects of these physiological processes on BOLD signal. In addition, the results suggest that whereas the neural component of the BOLD signal dimensionality is driven by interactions of network strength and connectivity, the dimensionality of global physiological noise is driven by intrinsic variability in respiration and heartbeat. However, the interactions of cardiac and respiratory variability are complex, as evidenced by the outliers in cardiac trends of Fig. 5. These high-dimensionality outliers suggest that the effects of respiratory and cardiac SD may uncouple in some cases, with respiration having a dominant effect on estimated physiological noise dimensionality; however, further research is required to establish any definitive conclusions.

We have also shown that the estimated structure and amplitude of physiological effect is highly dependent on task design. SART evidences lower-dimensional, more reproducible, higher Z-score signal effects, as compared with TMT. Given that this is performed for fixed subject group and scanning parameters, with comparable data acquisition times, the effect appears to be driven primarily by differences due to task design. We suggest that SART involves a more consistent level of physiological arousal, with lower respiratory and cardiac variability, since GO-NOGO “events” were driven by a consistent, rapid interstimulus interval, requiring relatively consistent cognitive engagement. This may be compared to the TMT design, in which “events” (tracing items) were subject-paced, and involved large changes in cognitive engagement, particularly between Tasks (A and B) and Control. It has been previously observed that regions of brain activation in TMT range from strong visuo-motor to primarily default-mode type patterns (Zakzanis *et al*., 2007; Tam *et al*., 2011; Churchill *et al*., *in press*), depending

on stimulus; this appears to be associated with more variable rates of physiological arousal. This is corroborated with our measurements of significant increases in cardiac and respiratory SD for TMT relative to SART; the increase in PHYCAA noise dimensionality for TMT thus appears to be related to increased physiological variability, and possibly interactions with the neuronally-coupled BOLD response. The gSNR of physiological noise may thus be an effective predictor of the extent (and type) of required physiological preprocessing, which is task-dependent. These findings are mirrored by results of (Churchill *et al*., *in submission*), which show that the extent of optimal physiological noise correction depends on the specific type of BOLD contrast, and based on these results potentially task design.

The results of PHYCAA denoising tend to show consistent improvements in signal detection and group-level statistics of regions with generally high activation. In addition, the removal of PHYCAA-derived noise has both localized and global effects, with signal increases that are distant from ventricles and vascular regions of greatest physiological noise amplitude. This is expected, given that both respiratory and cardiac fluctuations have been shown to induce signal change throughout gray matter regions (Birn *et al*., 2006; Shmueli *et al*., 2007), along with the well-known effects in ventricles and macrovasculature. More generally, these results are consistent with findings of Smith *et al*. (2007), who show that non-white physiological noise has a systematic bias effect throughout the brain.

Our results may also help inform general preprocessing strategy when correcting for the influence of physiological noise. As we have seen, a data-driven approach is generally preferable in the majority of subjects, since the primary physiological noise structure is not well specified by regressing external cardiac and respiratory measurements, except in a few subjects that improve with RETROICOR versus PHYCAA (Figs. 7 and 8). In addition, the CAA-based findings estimate the minimal spanning set of regressors required to minimize physiological noise effects. For relatively fixed task-conditions of SART, we see that approximately 2 regressors are required to optimize results, on average. Whereas for more designs with greater variation in levels of cognitive engagement (and presumably, physiological arousal), such as for TMT, we require 5 regressors on average, and thus more extensive modelling. This suggests that a fixed physiological noise preprocessing strategy is inadequate. The method of using single physiological regressors (e.g. white matter, CSF or large vessel seeds), as per Lund *et al*. (2001) and Petersen *et al*. (1998), would therefore not be expected to effectively denoise the TMT data.

One potential limitation of the current technique is that we restrict noise component selection to higher frequencies > 0.1 Hz. As previously shown (Lund et al., 2006), cardiac phase effects may be extensively aliased into the task-frequency range; in addition, the effects of change in cardiac rate and respiratory volume may be present in the task-frequency spectral band (Birn *et al*., 2006; Shmueli *et al*., 2007; Chang et al., 2009). The current model is conservative, as it is presently designed to remove the more directly separable high-frequency portion of physiological noise. However, the PHYCAA framework is flexible in design, which may allow us to substitute criteria that are more sensitive to physiological noise across the full spectral range. For example, we may deliberately undersample external cardiac and respiratory measurements, to identify the expected spectral bands of heavily aliased physiological artifact in fMRI data, along with slower effects of change in respiratory depth and cardiac rate. In addition, the algorithm may be modified to use alternative data types, such as complex-valued fMRI data. It has been shown that the capillary BOLD signal has distinct phase and magnitude characteristics, compared to large-vessel flow (Menon, 2002) and respiratory effects (Chen & Li, 2010); this information may be integrated as spatiotemporal priors to more precisely separate BOLD response from physiological artifact, and/or the whole analysis could be readily recast into a framework using complex multivariate techniques (Rowe & Logan, 2004).

In recent years, a number of alternate data-driven models have also been developed, with the goal of estimating and removing physiological noise (Behzadi et al., 2007; Perlbarg *et al*., 2007; Tohka *et al*., 2008; Beall et al., 2010a,b). We suggest that the proposed PHYCAA model provides advantages in that it is built on a set of statistically rigorous, data-driven methods, requiring no qualitative estimation of spatial priors or spectral peaks. In addition, we have shown that identification of physiological noise requires adaptive subspace estimation, and that the reproducibility measures used can optimize subspace estimation of correlated, spatially-distributed BOLD signals (Yourganov et al., 2011). To our knowledge, no pre-existing method optimizes physiological noise estimation in such a quantitative resampling framework. By comparison, other techniques often perform estimation without using power-spectrum constraints (e.g. Behzadi et al., 2007; Perlbarg *et al*., 2007; Tohka *et al*., 2008), and certain cases, take advantage of lags in slice acquisition for increased temporal sensitivity (Beall et al., 2010a). It

remains to be investigated which of these features of the various physiological noise estimators are of greatest individual and combined importance.

It is thus important to compare PHYCAA with preexisting data-driven models in a consistent resampling framework, to determine which methods optimize model prediction and reproducibility. Of particular interest is the denoising efficacy for resting-state analyses, which involve techniques that are often highly sensitive to the high-variance, high autocorrelation of physiological noise, and lack an overt task-design for predictive model validation (see Cole et al., 2010 for an overview of these issues). We have performed the current analyses using overt task designs in order to validate PHYCAA's effects, including predictive accuracy. However, in this paper we have examined two models that involve asynchronous, relatively weak cognitive contrasts, and used multivariate/locally-multivariate analysis models which are also sensitive to the high-variance effects of physiological noise, given their dependence on the underlying data covariance structure (Mardia et al., 1979; Strother et al., 2010). This provides some evidence of PHYCAA's potential efficacy in denoising, although direct comparison against other available techniques in a resting-state framework must be addressed in future work.

5. APPENDIX: Dimensionality and Degeneracy in CCA

CCA is applied to matrices $\boldsymbol{X}_{kxN}$ and $\boldsymbol{Y}_{kxN}$ (*N* samples in k-dim. space). The decomposition is specified by nominally independent basis sets $\boldsymbol{C_x}_{(kxk)}$ and $\boldsymbol{C_y}_{(kxk)}$ (of ***X*** and ***Y*** respectively), and diagonal correlation matrix $\boldsymbol{R}_{kxk}$. The representation has net degrees of freedom (*DOF*):

$$DOF = dof(U) + dof(V) + dof(R)$$

$$DOF = (k \times k) + (k \times k) + k = 2k^2 + k \qquad \text{[A.1]}$$

To obtain a non-degenerate solution for [A.1], we require sufficient independent parameters of ***X*** and ***Y*** to equal the *DOF*. Since ***X*** and ***Y*** each have *nk* components, we obtain the inequality $2nk \geq 2k^2 + k$.

For a fixed sample size *n*, a nondegenerate CCA solution imposes the dimensionality limit:

$$k \leq (2n-1)/2 \approx n \qquad \text{[A.2]}$$

That is, dimensionality must not exceed the number of total sample points. However, under Canonical Autocorrelations Analysis (CAA) and 1 TR timeshift, ***X*** and ***Y*** are now of size k·(n-1). Furthermore, n-2 timepoints are shared between matrices, with only 1 timepoint unique to each. The number of independent parameters is now $k(n-1) + 2k(1) = kn$. Thus the new inequality, assuming $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$ remain approximately independent, of $nk \geq 2k^2 + k$. This provides restriction on dimensionality *k* of:

$$k \leq (n-1)/2 \approx n/2 \qquad \text{[A.3]}$$

Under these conditions, we require that dimensionality *k* not exceed 50% of sample size, to provide a well-conditioned (e.g. non-degenerate) solution. It should be noted that [A.3] is the most conservative dimensionality limit; excluding edge effects, the paired, autocorrelated vectors of $\boldsymbol{C}_x$ and $\boldsymbol{C}_y$ are likely to be intrinsically generated by similar spatial regions, and/or PC bases. This would potentially reduce the *DOF* of the CAA decomposition, in turn reducing the upper limit on dimensionality *k*.

6. REFERENCES

Afshin-Pour B, Hossein-Zadeh GA, Soltanian-Zadeh H, Grady C, Strother SC. (2010). Enhancing the reproducibility of fMRI statistical maps using generalized canonical correlation analysis in the NPAIRS framework. OHBM, 2010

Beall EB. (2010**a**). Adaptive cyclic physiologic noise modeling and correction in functional MRI. J Neurosci Methods 187(2):216-228.

Beall EB, Lowe MJ. (2010**b**): The non-separability of physiologic noise in functional connectivity MRI with spatial ICA at 3T. J Neurosci Methods 191(2):263-76.

Beckmann CF, Noble JA and Smith SM. (2004). Artefact detection in FMRI data using independent component analysis. NeuroImage 11(5): S614

Behzadi Y, Restom K, Liau J, Liu TT. (2007). A Component Based Noise Correction Method (CompCor) for BOLD and Perfusion Based fMRI. NeuroImage 37(1):90-101

Birn RM, Diamond JB, Smith MA, and Bandettini PA. (2006). Separating respiratory-variation-related fluctuations from neuronal-activity-related fluctuations in fMRI. NeuroImage 31:1536-1548

Birn RM, Murphy K, Handwerker DA, Bandettini PA. (2010). fMRI in the presence of task-correlated breathing variations. NeuroImage 47(3):1092-1104

Biswal B, Yetkin FZ, Haughton VM, Hyde JS. (1995). Functional Connectivity in the Motor Cortex of Resting Human Brain Using Echo-Planar MRI. MRM 34:537-541

Bollen KA, Jackman R. (1990). Regression diagnostics: An expository treatment of outliers and influential cases. In: J. Fox & J. Scott Long (eds.) Modern Methods of Data Analysis (pp. 257-91). Newbury Park: Sage.

Bullmore ET, Rabe-Hesketh S, Morris RG, Williams SCR, Gregory L, Gray JA, Brammer MJ. (1996). Functional Magnetic Resonance Image Analysis of a Large-Scale Neurocognitive Network. NeuroImage 4:16–33

Chang C, Cunningham JP, Glover GH. (2009). Influence of heart rate on the BOLD signal: The cardiac response function. NeuroImage 44:857-869

Cheng H, Li Y. (2010). Respiratory noise correction using phase information. MRM 28(4):574-582

Chen EE & Small SE. (2007). Test–retest reliability in fMRI of language: Group and task effects. Brain and Language. 102(2): 176-185

Chuang KH, Chen JH. (2001). IMPACT: Image-based physiological artifacts estimation and correction technique for functional MRI. Magn Reson Med 46(2):344-353.

Churchill N, Oder A, Abdi H, Tam F, Lee W, Thomas C, Ween J, Graham S, Strother S. (in press). Optimizing preprocessing and analysis pipelines for singlesubject fMRI: 1. Standard temporal motion and physiological noise correction methods. HBM (in press)

Churchill NW, Yourganov G, Oder A, Tam F, Graham SJ,Strother SC (in submission). Optimizing preprocessing and analysis pipelines for single-subject fMRI: 2. The Effects of Analysis Model and Task Contrast. NeuroImage

Cole DM, Smith SM, Beckmann CF. (2010). Advances and pitfalls in the analysis and interpretation of

resting-state FMRI data. Front Syst Neurosci 4(8):1-15

Cook RD, Weisber S.(1982). Residuals and influence in regression. New York: Chapman & Hall.

Cox, RW (1996). AFNI: software for analysis and visualization of functional magnetic resonance neuroimages. Comput Biomed Res 29:162–173

Dagli MS, Ingeholm JE, Haxby JV.(1999). Localization of cardiac-induced signal change in fMRI. Neuroimage.9(4):407–415.

De Martino F, Gentile F, Esposito F, Balsi M, Di Salle F, Goebel R, Formisano E. (2007). Classification of fMRI independent components using IC-fingerprints and support vector machine classifiers, NeuroImage 34(1):177-194

Fassbender C, Murphy K, Foxe JJ, Wylie GR, Javitt DC, Robertson IH, Garavan H.(2004). A topography of executive functions and their interactions revealed by functional magnetic resonance imaging, Brain Res Cogn Brain Res*,* 20:132-43.

Friman O, Borga M, Lundberg P, Knutsson H. (2002). Exploratory fMRI Analysis by Autocorrelation Maximization. NeuroImage 16, 454–464

Foster PS, Harrison DW. (2004). The covariation of cortical electrical activity and cardiovascular responding. Int J Psychophysiol. 52(3):239-255

Glover GH, Li TQ, and Ress D. (2000). Image-Based Method for Retrospective Correction of Physiological Motion Effects in fMRI: RETROICOR. Magnetic Resonance in Medicine 44:162-167

Guimond, A., Meunier, J., Thirion, J. (2000). Average brain models: A convergence study. Computer Vision and Image Understanding *77*, 192-210.

Hansen LK, Larsen J, Nielsen FA, Strother SC, Rostrup E, Savoy R, Lange N, Sidtis J, Svarer C, Paulson, OB. (1999). Generalizable patterns in neuroimaging: how many principal components? Neuroimage *9*, 534-544.

Hilton M, Ogden T, Hattery D, Eden G, Jawerth B. (1996). Wavelet Denoising of Functional MRI Data. Wavelets in Medicine and Biology, Aldroubi, A. & Unser, M. (eds.), CRC Press: Washington D.C., pp. 93-114

Hotelling H. (1935). The Most Predictable Criterion. Journal of Educational Psychology. 26:139-142

Hu X, Le TH, Parrish T, Erhard P. (1995). Retrospective estimation and correction of physiological fluctuation in functional MRI. MRM 34(2):201-212

Jahanian H, Soltanian-Zadeh H, and Hossein-Zadeh GA. (2005). Noise Suppression of fMRI Time-Series in Wavelet Domain. Proceedings from Signal and Image Processing. pp 479.

Jones TB, Bandettini PA, and Birn RM. (2008). Integration of motion correction and physiological noise regression in fMRI. NeuroImage, 42(2):582-590

Kelly RE, Alexopoulos GS, Wang Z, Gunning FM, Murphy CF, Morimoto SS, Kanellopoulos D, Jia Z, Lim KO and Hoptman MJ. (2010). Visual inspection of independent components: Defining a procedure for artifact removal from fMRI data. Journal of Neuroscience Methods 189(2):233-245

Kettenring, J. (1971). Canonical analysis of several sets of variables. Biometrika 58:433-451.

Kiviniemi V, Kantola JH, Jauhiainen J, Hyvarinen A, Tervonena O. (2003). Independent component

analysis of nondeterministic fMRI signal sources. NeuroImage 19:253-260

Kjems U, Hansen LK, Anderson J, Frutiger S, Muley S, Sidtis J, Rottenberg D, Strother SC (2002). The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. Neuroimage 15:772–786.

Kriegeskorte N, Goebel R, Bandettini P. (2006). Information-based functional brain mapping. Proc Natl Acad Sci USA 103(10):3863-8.

Lukic AS, Wernkick MN, Hansen LK, Anderson J, Strother SC. (2002). A spatially robust ICA algorithm for multiple fMRI data sets. In IEEE International Symposium on Biomedical Imaging, Proceedings, pp 839–842.

Lukic AS, Wernick MN, Strother SC. (2002). An evaluation of methods for detecting brain activations from functional neuroimages. Artificial Intelligence in Medicine, 25(1):69-88

Lund, TE, Hanson, LG (2001). Physiological noise correction in fMRI using vessel time-series as covariates in a general linear model. In: Proceedings of the 9th Annual Meeting of ISMRM, pp 22.

Lund TE, Madsen KH, Sidaros K, Luo WL, Nichols TE (2006). Non-white noise in fMRI: Does modelling have an impact? NeuroImage 29(1):54-66

Mardia K, Kent J, Bibby J. (1979). Multivariate analysis. Academic Press. London, United Kingdom

McKeown MJ, Makeig S, Brown GG, Jung TP, Kindermann SS, Bell AJ, Sejnowski TJ. (1998). Analysis of fMRI data by blind separation into independent spatial components. Human Brain Mapping, 6(3):160–188

Menon RS. (2002). Postacquisition Suppression of Large-Vessel BOLD Signals in High-Resolution fMRI. MRM 47:1-9

Moldegy L & Schuster HG. (1994). Separation of a Mixture of Independent Signals Using Time Delayed Correlations. Physical Review Letters. 72(23): 3634-3737

Penny W, Kiebel S, Friston K. (2003). Variational Bayesian inference for fMRI time series. NeuroImage 19:727–741.

Pereira F, Mitchell T, Botvinick M. (2009). Machine learning classifiers and fMRI: a tutorial overview. Neuroimage 45(1 Suppl): S199–S209

Perlbarg V, Bellec P, Anton JL, Pélégrini-Issac M, Doyon J, Benali H. (2007). CORSICA: correction of structured noise in fMRI by automatic identification of ICA components. Magn Reson Imaging. 25(1):35-46.

Petersen, NV, Jensen JL, Burchardt J, Stødkilde-Jørgensen H. (1998). State space models for physiological noise in fMRI time series. NeuroImage,7(4), s592.

Petersen KS, Hansen LK, Kolenda T, Rostrup E, Strother SC. (2000). On the dependent components of functional neuroimages. In: Proc. ICA-2000, (Pajunen P, Karhunen J, Eds.), pp. 615-620, Helsinki, Finland.

Rowe DB, Logan BR. (2004). A complex way to compute fMRI activation. NeuroImage, 23(3):1078-1092

Shmueli K, vanGelderen P, deZwart JA, Horovitz SG, Fukunaga M, Jansma JM, Duyn JH. (2007). Low-frequency fluctuations in the cardiac rate as a source of variance in the resting-state fMRI BOLD signal. NeuroImage 38(2):306-320

Song X, Murphy M, Wyrwicz AM. (2006). Spatiotemporal Denoising and Clustering of fMRI Data. Image Processing, 2006 IEEE International Conference on. 2857 - 2860

Song X, Ji T, Wyrwicz AM. (2008). Baseline drift and physiological noise removal in high field FMRI data using kernel PCA. Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on. pp. 441-444

Smith SM, Jenkinson,M,Woolrich MW, Beckmann CF, Behrens TEJ, Johansen-Berg H, Bannister PR, De Luca M, Drobnjak I, Flitney DE, Niazy R, Saunders J, Vickers J, Zhang Y, De Stefano N, Brady JM, Matthews PM. (2004). Advances in functional and structural MR image analysis and implementation as FSL. NeuroImage 23 (S1), 208–219 [http://www.fmrib.ox.ac.uk/fsl/].

Smith AT, Singh KD, Balsters JH. (2007). A comment on the severity of the effects of non-white noise in fMRI time-series. NeuroImage 36:282-288

Strother SC, Anderson J, Hansen LK, Kjems U, Kustra R,Sidtis J, Frutiger S, Muley S, LaConte S and Rottenberg D. (2002). The Quantitative Evaluation of Functional Neuroimaging Experiments: The NPAIRS Data Analysis Framework. NeuroImage 15:747–771

Strother S, Oder A, Spring R, Grady C. (2010): The NPAIRS Computational Statistics Framework for Data Analysis in Neuroimaging. In: Saporta G, et. al., editor. 19th International Conference on Computational Statistics. Paris. Physica-Verlag, Heidelberg.

Tam F, Churchill NW, Strother SC, Graham SJ. (2011). A new tablet for writing and drawing during functional MRI. Hum Brain Mapp 32(2):240-8

Taylor HL, Haskell W, Fox SM III, Blackburn H.(1969). Exercise tests: a summary of procedures and concepts of stress testing for cardiovascular diagnosis and function evaluation. Measurement in Exercise Electrocardiography. Ernst Simonson Conference, H Blackburn (ed). Springfield, Illinois.

Tohka J,Foerde K,Aron AR, Tom SM, Toga AW and Poldrack RA. (2008). Automatic Independent Component Labeling for Artifact Removal in fMRI. Neuroimage. 2008 February 1; 39(3): 1227–1245.

Wang Z & Peterson BS. (2008). Partner-Matching for the Automated Identification of Reproducible ICA Components from fMRI Datasets: Algorithm and Validation. HBM 29:875–893

Windischberger C, Langenberger H, Sycha T, Tschernko EM, Fuchsjager-Mayerl G, Schmetterer L, Moser E.(2002). On the origin of respiratory artifacts in BOLD-EPI of the human brain. *Magn Reson Imaging*. 20:575–82.

Windischberger C, Barth M, Lamm C, Schroeder L, Bauer H, Gur RC, Moser E. (2003). Fuzzy cluster analysis of high-field functional MRI data . Artificial intelligence in medicine (Tecklenburg, Germany), 29(3): 203-223

Worsley KJ. (2001). Statistical analysis of activation images. In: Jezzard P, Matthews PM, Smith SM. (Eds.), Functional MRI: An Introduction to Methods. Oxford University Press, NY, pp. 251–270.

Yang Z, LaConte S, Weng X, Hu X. (2008). Ranking and averaging independent component analysis by reproducibility (RAICAR). HBM. 29(6): 711–725

Yourganov G, Chen X, Lukic AS, Grady CL, Small SL, Wernick MN, Strother SC. (201). Dimensionality

estimation for optimal detection of functional networks in BOLD fMRI data. Neuroimage, 56(2):531-43

Zollei L, Panych L, Grimson E, Wells WM. (2003). Exploratory identification of cardiac noise in fMRI images. Medical Image Computing and Computer-Assisted Intervention—MICCAI 2003, Pt. 1, vol. 2878, Lecture Notes in Computer Science, pp. 475–482

Zou QH, Zhu CZ, Yang Y, Zuo XN, Long XY, Cao QJ, Wang YF, Zang YF. (2008). An improved approach to detection of amplitude of low-frequency fluctuation (ALFF) for resting-state fMRI: fractional ALFF. Journal of Neuroscience Methods 172(1):137-141

TABLE CAPTIONS

Table I: commonly used in-text abbreviations

Table II: correlations between optimal PC dimensionality (maximizing spatial reproducibility for PHYCAA), and externally-measured physiological processes, of respiratory frequency, variability in respiratory amplitude, and cardiac rate. The 95% confidence intervals are shown on Spearman correlations, obtained via Bootstrap resampling (1000 iterations). Probabilities are given by the fraction of Bootstrap resamples with correlations > 0. Significant correlations, defined by 95% confidence intervals that do not include zero, are in bold.

FIGURE CAPTIONS

Figure 1: schematic of PHYCAA algorithm used to estimate and remove physiological noise from fMRI data. The fMRI data is represented as split-half 2D matrices, of dimensions (**n** voxels × **t** timepoints), with maximum principal component dimensionality $\mathbf{K} \leq \mathbf{t}$.

Figure 2: matrix condition of cross-covariance S_{xy} (used in Eqtn. [3,4]) as a function of PC dimensionality, displayed on logarithmic scale. Results are plotted for all subjects, for the Trail-Making Test; note that PC dimensionalities $k > 70$ cannot be displayed at this scale, due to extremely high condition numbers.

Figure 3: (a,b) intensity maps showing fraction of power spectrum in physiological noise frequency range (>0.10 Hz) for each Canonical Variate (CV), as a function of PC dimensionality. Each pixel represents the mean fraction, computed across all subjects. CVs are ordered by decreasing autocorrelation significance, and the black curve on the plot marks the median p=0.05 significance cutoff for each PC dimensionality; CVs below this cutoff tend to be discarded as nonsignificant. (c,d) distributions of the number of significant CVs (p=0.05) with > 50% of spectral power above 0.10 Hz, as a function of PC dimensionality; the blue dashed line represents the identity line, at which (# significant CVs = PC dimensionality). For all plots, the t/2 theoretical dimensionality limit is plotted as a vertical black line. Results are shown for Trail-Making Test, and Sustained Attention to Response Task experimental designs.

Figure 4: PC dimension as a function of global signal-to-noise, estimated from physiological effect maps of the PHYCAA algorithm. Results are plotted for the PC dimensionality optimizing spatial reproducibility, as determined on an individual-subject basis. Values are plotted for both Trail-Making Test (TMT) and Sustained Attention to Response Task (SART) experimental designs. Note that SART produces a high-signal datapoint that cannot be displayed at this scale.

Figure 5: plots of the relation between optimized PC dimensionality (e.g. maximizing spatial reproducibility), and the standard deviation (SD) of respiratory rate (a,d), respiratory amplitude (b,e) and cardiac rate (c,f). Results (a-c) are shown for Trail-Making Test (TMT), and results (d-f) are shown for Sustained Attention to Response Task (SART). High-dimensionality outliers of the cardiac SD trends are coloured in white. The estimated 2^{nd}-order polynomial of best fit is plotted (solid line), with 95% confidence range, estimated via Bootstrap resampling (1000 iterations) are also shown (dashed lines). High dimensionality outliers are not included for the cardiac rate curve fitting.

Figure 6: spatial maps of brain regions with greatest estimated physiological noise variance using the PHYCAA algorithm. The plots show mean Z-scored pseudo-F maps, as computed across all subjects on a per-voxel basis, for (a) Trail-Making Test (TMT) and (b) Sustained Attention to Response Task (SART).

Figure 7: effects of PHYCAA denoising on simulated data, with added respiratory and cardiac artifact. (a) spatial maps showing loci of task activation (left) cardiac-driven artifact (middle) and respiratory-driven artifact (right). (b) Receiver Operating Characteristic (ROC) curve, comparing True Positive Rates (detection in activation loci) to False Positive Rates (detection in null pixels; e.g. regions without task or artifact sources). (c-d) rate of false positives in artifact loci, as a function of Z-score threshold for cardiac and respiratory sources.

Figure 8: effects of PHYCAA denoising on analysis results for multivariate Penalized Discriminant Analysis, on the Trail-Making Test. Changes in NPAIRS metrics of (a) reproducibility and (b) prediction on standard preprocessing results, by both PHYCAA and the commonly-used RETROICOR. Subjects with improved performance relative to basic preprocessing (no physiological correction) are plotted in blue,

and decreased performance plotted in red. (b) Maps showing spatial effect of PHYCAA regression. The voxelwise mean Z-scored SPM is plotted for both basic preprocessing and PHYCAA denoising, along with the mean Z-score changes due to denoising; the latter is thresholded for significant change using Bootstrap estimates (1000 iterations), corrected for multiple comparisons at FDR=0.10.

Figure 9: effects of PHYCAA denoising on analysis results for searchlight GNB, for the Sustained Attention to Response Task. (a) percent change in significant voxels is shown as a function of prediction threshold, for PHYCAA and RETROICOR denoising, relative to basic preprocessing (no physiological correction); each point represents a single-subject result. (b) Effects on the spatial prediction maps are shown, including mean voxel-wise prediction accuracy, computed across all subjects, and regions with consistent change in prediction accuracy, with voxelwise significance by Bootstrap estimation (1000 iterations), thresholded at FDR=0.10.

4. Table 1

PHYCAA	PHYsiological correction using Canonical Autocorrelation Analysis
RETROICOR	Retrospective physiological correction based on respiratory and cardiac phase
PCA	principal component analysis
CAA	canonical autocorrelation analysis
CV	canonical variates (CAA timeseries)
TMT	Trail-Making Test
SART	Sustained Attention to Response Task
PDA	Penalized Discriminant Analysis
slGNB	searchlight Gaussian Naïve Bayes
gSNR	global Signal-to-Noise Ratio
R	reproducibility (spatial)
P	prediction accuracy (temporal)
CI	confidence interval

4. Table 2

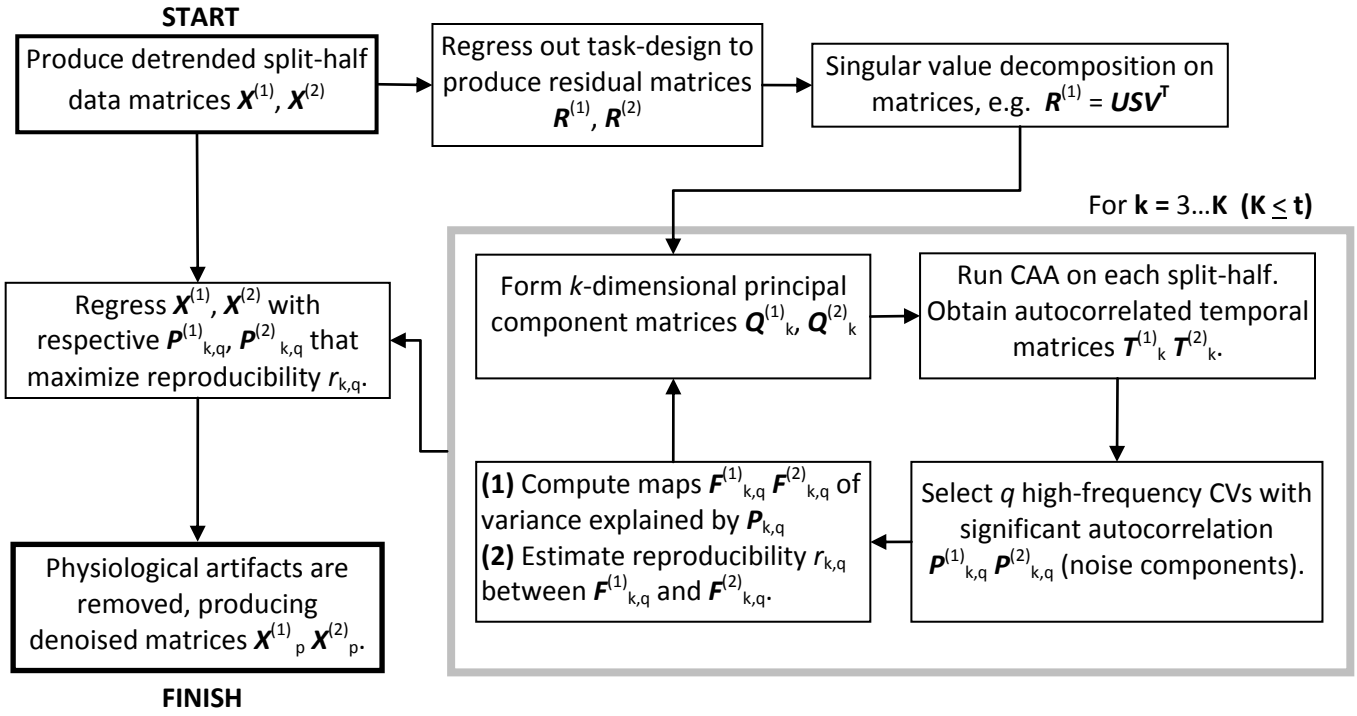
(a) TMT		Spearman correlation	95 % confidence interval		probability of correlation > 0
Respiratory frequency (Hz)	mean	-0.11	-0.57	0.37	0.658
	SD	**0.47**	**0.07**	**0.77**	**0.041**
Respiratory amplitude (% of range)	mean	--	--	--	--
	SD	**0.49**	**0.06**	**0.83**	**0.034**
Cardiac frequency (Hz) *	mean	0.28	-0.29	0.76	0.27
	SD	**0.69**	**0.15**	**0.92**	**0.002**

(b) SART		Spearman correlation	95 % confidence interval		probability of correlation > 0
Respiratory frequency (Hz)	mean	-0.06	-0.52	0.41	0.822
	SD	**0.60**	**0.19**	**0.90**	**0.006**
Respiratory amplitude (% of range)	mean	--	--	--	--
	SD	**0.61**	**0.18**	**0.89**	**0.005**
Cardiac frequency (Hz) *	mean	0.35	-0.35	0.77	0.163
	SD	**0.58**	**0.18**	**0.77**	**0.028**

*cardiac results shown with high-dimensionality outliers removed (see Fig.5). For correlations with outliers included, see **3.2 Physiological Noise Reproducibility and PC-space Optimization**.

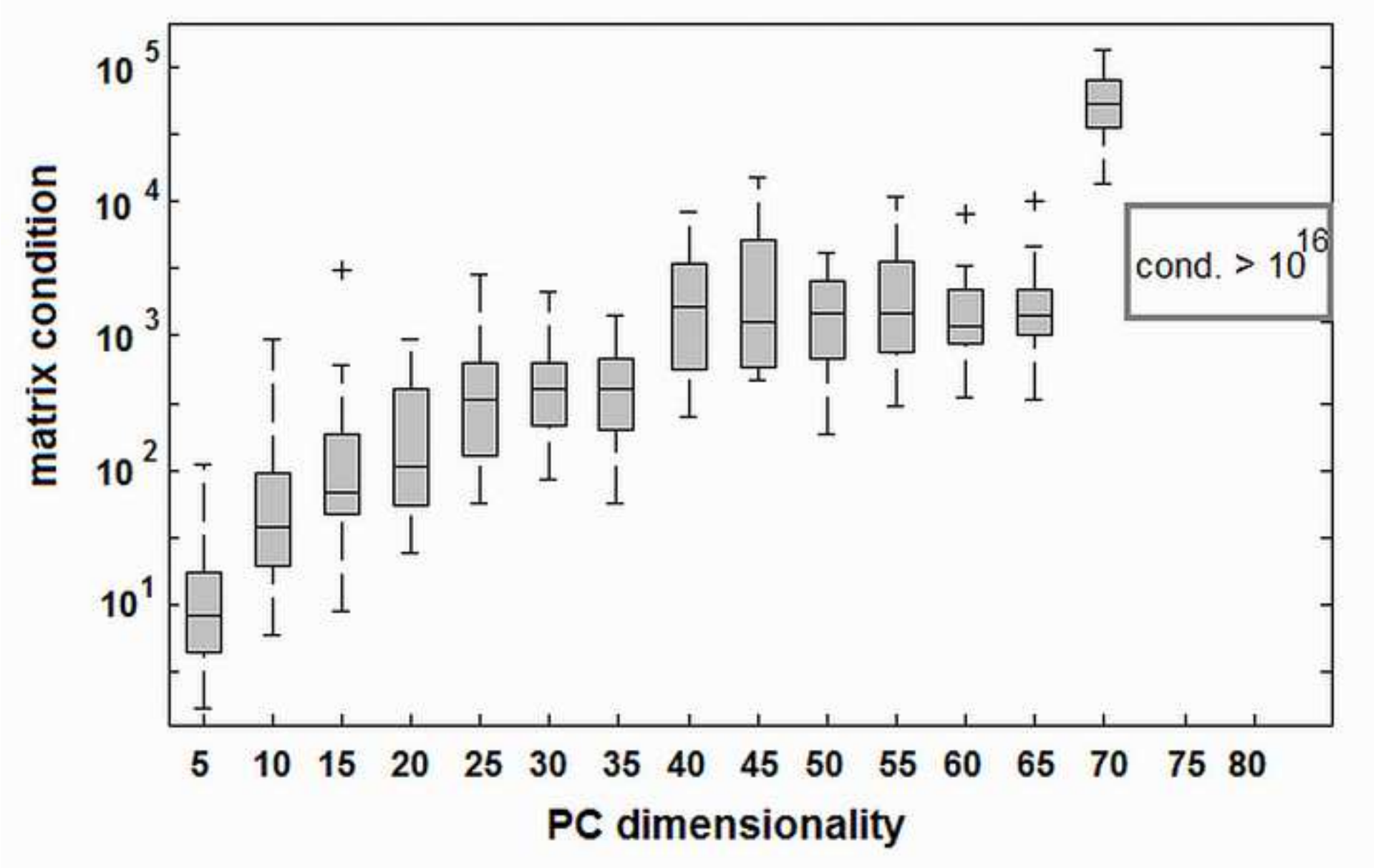
5. Figure 1

Click here to download 9. Figure: figure_1.docx



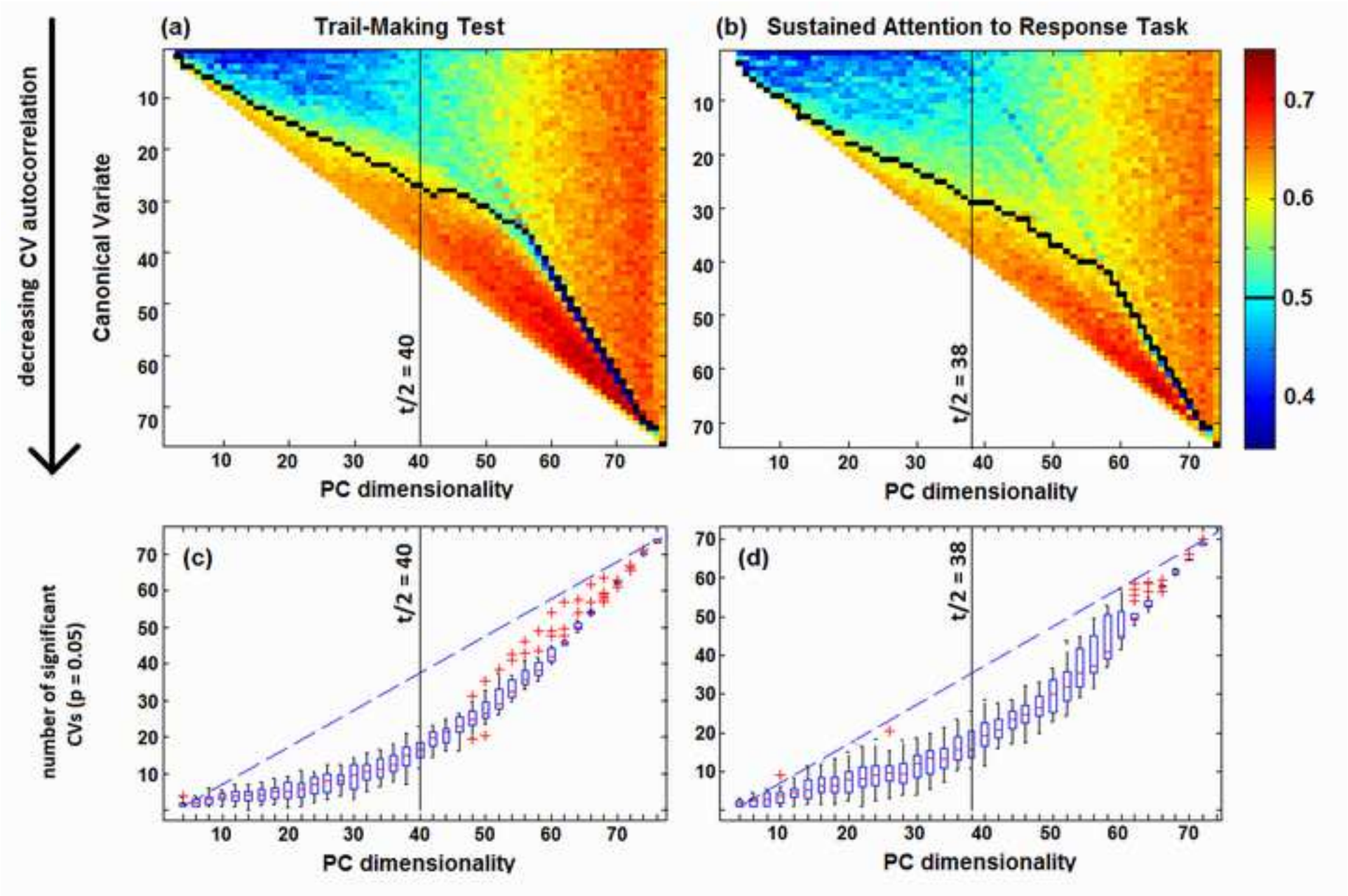
5. Figure 2

Click here to download high resolution image



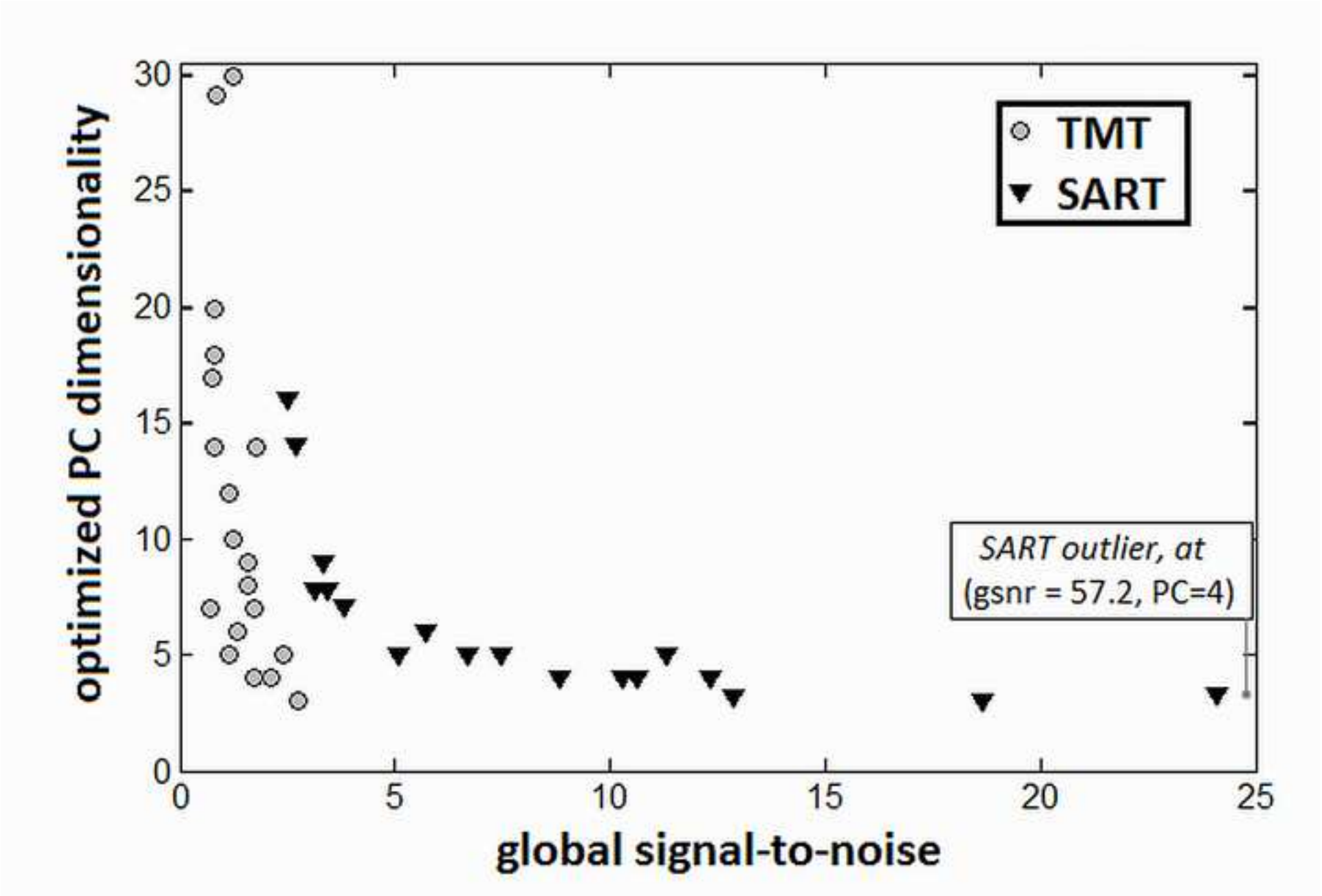
9. Figure 3

Click here to download high resolution image



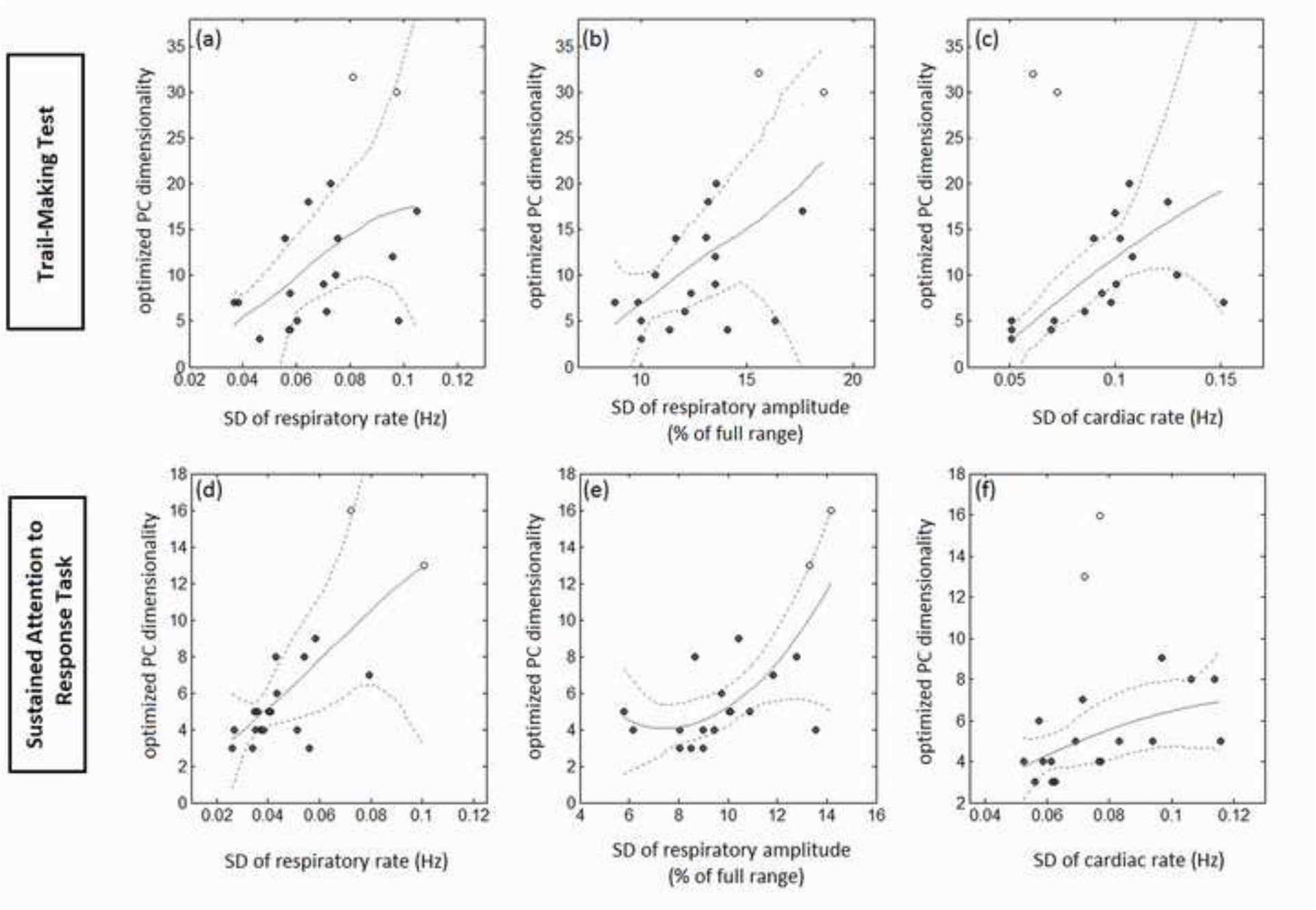
5. Figure 4

Click here to download high resolution image



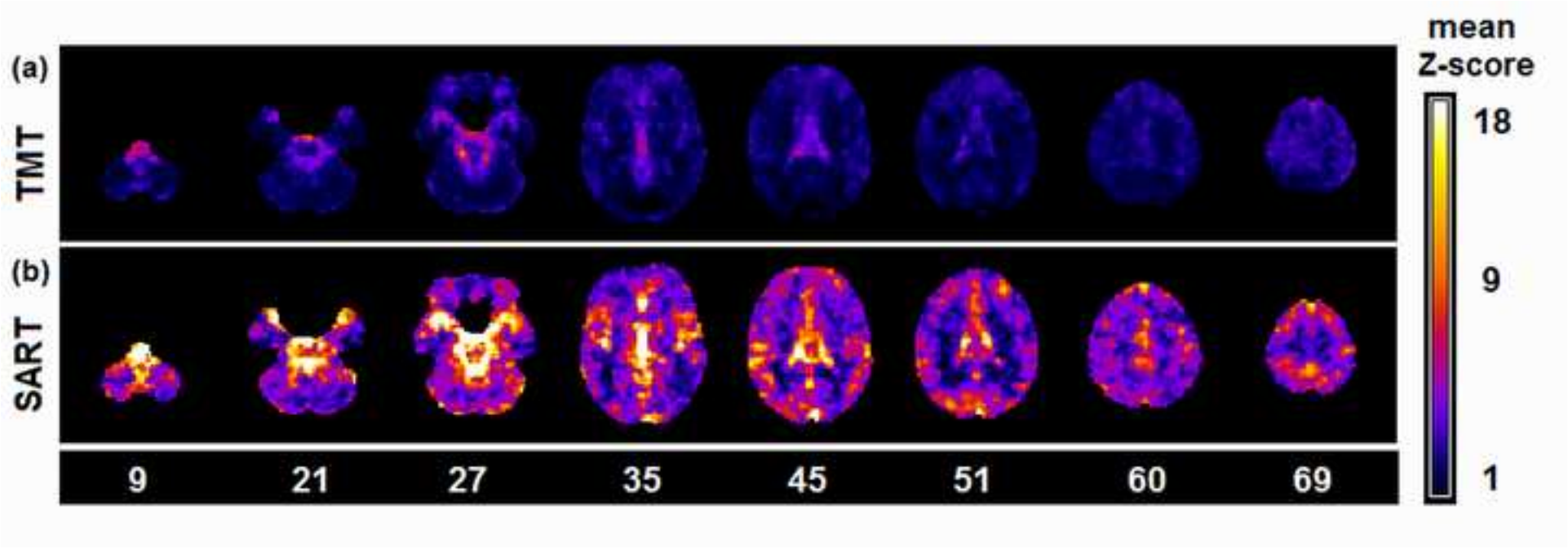
5. Figure 5

Click here to download high resolution image



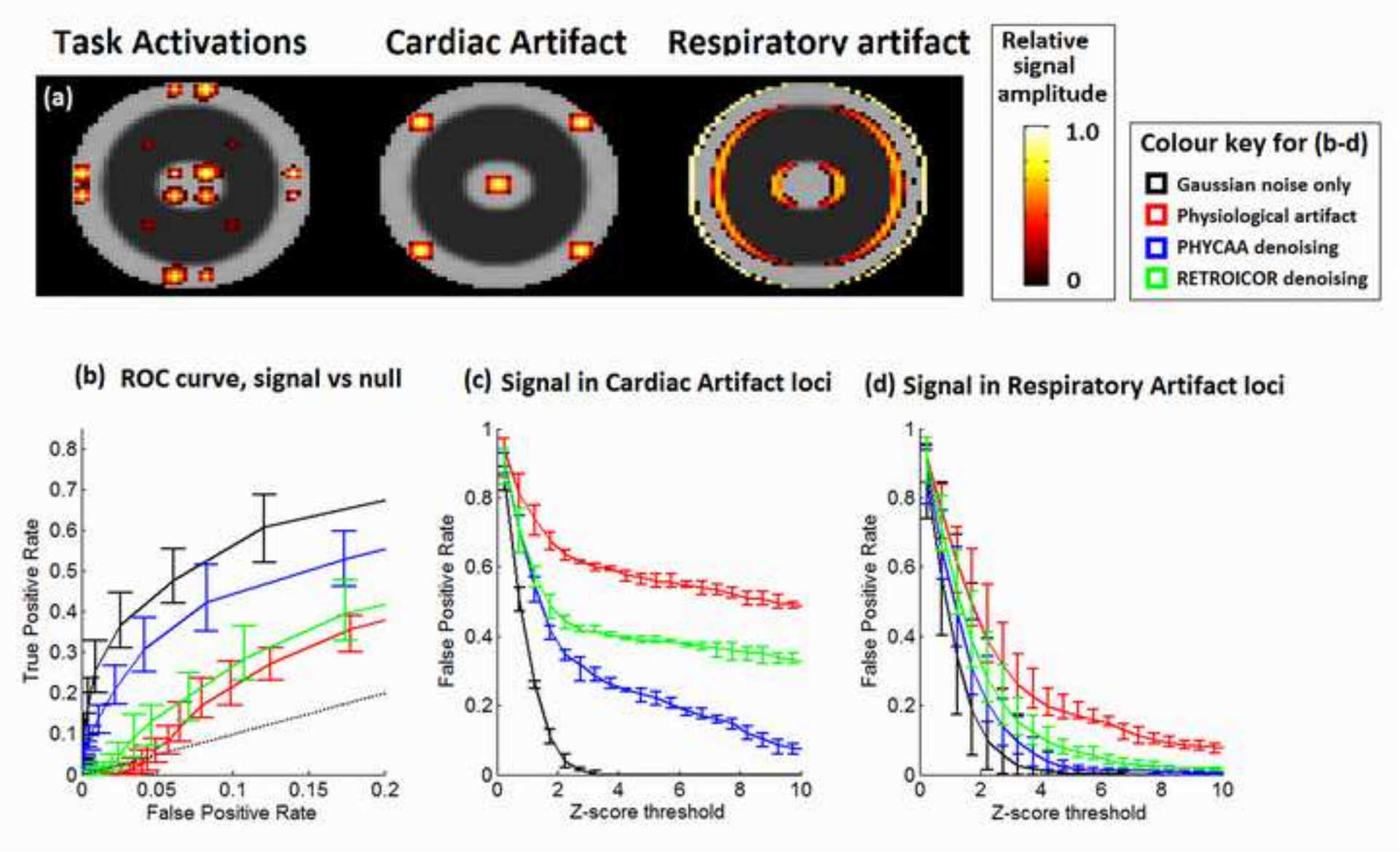
5. Figure 6

Click here to download high resolution image



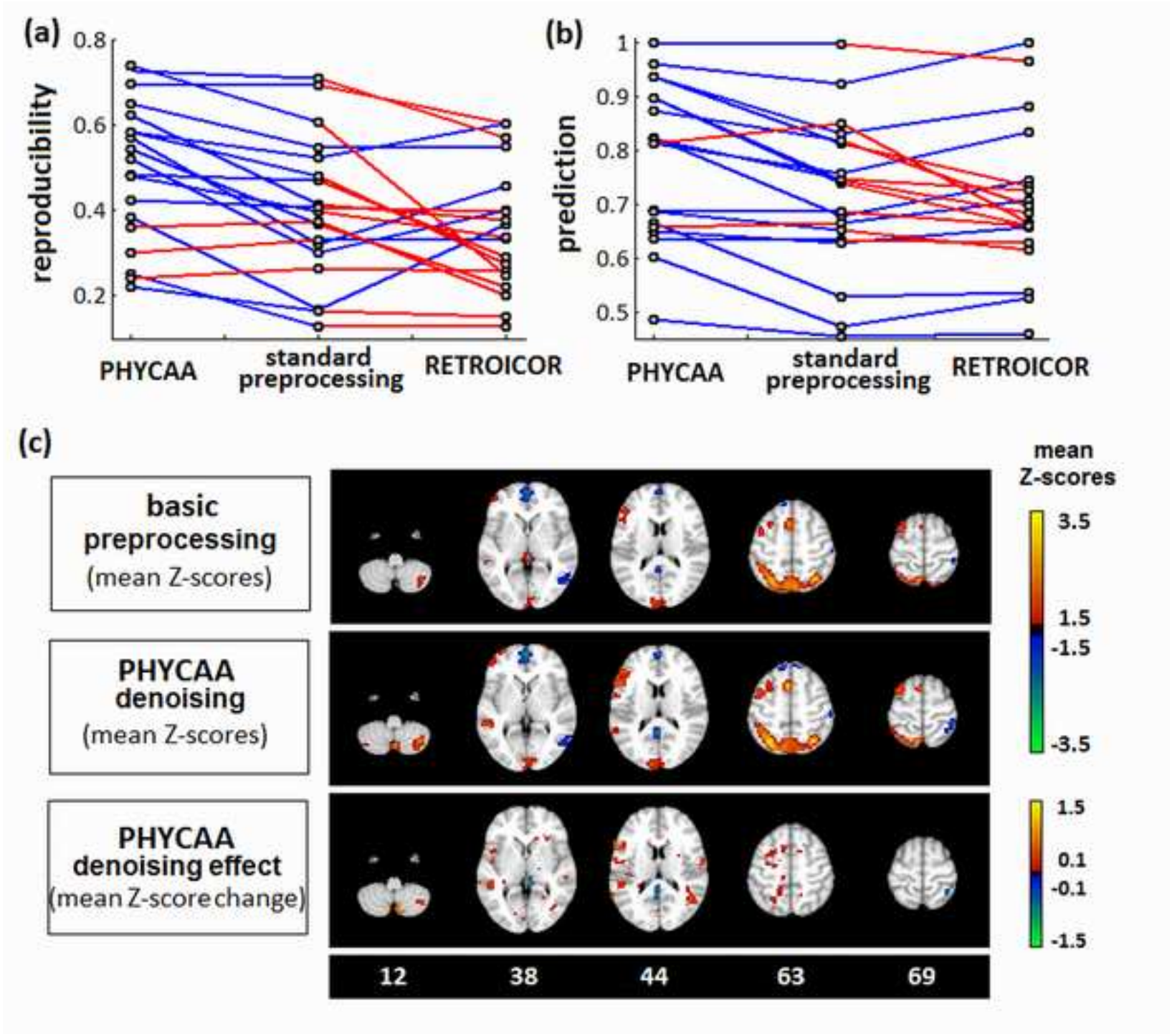
9. Figure 7

Click here to download high resolution image



9. Figure 8

Click here to download high resolution image



9. Figure 9

Click here to download high resolution image