REPRESENTATION LEARNING FOR IMPROVED INTERPRETABILITY AND CLASSIFICATION ACCURACY OF CLINICAL FACTORS FROM EEG

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite extensive standardization, diagnostic interviews for mental health disorders encompass substantial subjective judgment. Previous studies have demonstrated that EEG-based neural measures can function as reliable objective correlates of depression, or even predictors of depression and its course. However, their clinical utility has not been fully realized because of 1) the lack of automated ways to deal with the inherent noise associated with EEG data at scale, and 2) the lack of knowledge of which aspects of the EEG signal may be markers of a clinical disorder. Here we adapt an unsupervised pipeline from the recent deep representation learning literature to address these problems by 1) learning a disentangled representation using β-VAE to denoise the signal, and 2) extracting interpretable features associated with a sparse set of clinical labels using a Symbol–Concept Association Network (SCAN). We demonstrate that our method is able to outperform the canonical hand-engineered baseline classification method on a number of factors, including participant age and depression diagnosis. Furthermore, our method recovers a representation that can be used to automatically extract denoised Event Related Potentials (ERPs) from novel, single EEG trajectories, and supports fast supervised re-mapping to various clinical labels, allowing clinicians to re-use a single EEG representation regardless of updates to the standardized diagnostic system. Finally, single factors of the learned disentangled representations often correspond to meaningful markers of clinical factors, as automatically detected by SCAN, allowing for human interpretability and post-hoc expert analysis of the recommendations made by the model.

1 INTRODUCTION

Mental health disorders make up one of the main causes of the overall disease burden worldwide (Vos et al., 2013), with depression (e.g., Major Depressive Disorder, MDD) believed to be the second leading cause of disability (Lozano et al., 2013; Whiteford et al., 2013), and around 17% of the population experiencing its symptoms at some point throughout their lifetime (McManus et al., 2016; 2009; Kessler et al., 1993; Lim et al., 2018). At the same time diagnosing mental health disorders has many well-identified limitations (Insel et al., 2010). Despite the existence of diagnostic manuals like Structured Clinical Interview for Diagnostic and Statistical Manual of Mental Disorders (SCID) (DSM-V, 2013), diagnostic consistency between expert psychiatrists and psychologists with decades of professional training can be low, resulting in different diagnoses in upwards of 30% of the cases (Cohen's Kappa = 0.66) (Lobbestael et al., 2011). Even if higher inter-rater reliability was achieved, many psychological disorders do not have a fixed symptom profile, with depression alone having many hundreds of possible symptom combinations (Fried & Nesse, 2015). This means that any two people with the same SCID diagnosis can exhibit entirely different symptom expressions. This is a core challenge for developing an objective, symptom-driven diagnostic tool in this domain.

Electroencephalography (EEG) is a measurement of post-synaptic electrical potentials that can be taken non-invasively at the scalp. EEG signals can function as important biomarkers of clinical disorders (Hajcak et al., 2019) but they are difficult to clean and interpret at scale. For example, components of the EEG signal can often significantly overlap or interfere with each other. Furthermore, nearby electronics, line noise, hardware quality, signal drift and other variations in the electrode–scalp connection can all distort the recorded EEG signal. Hence, the extraction of EEG data of sufficient

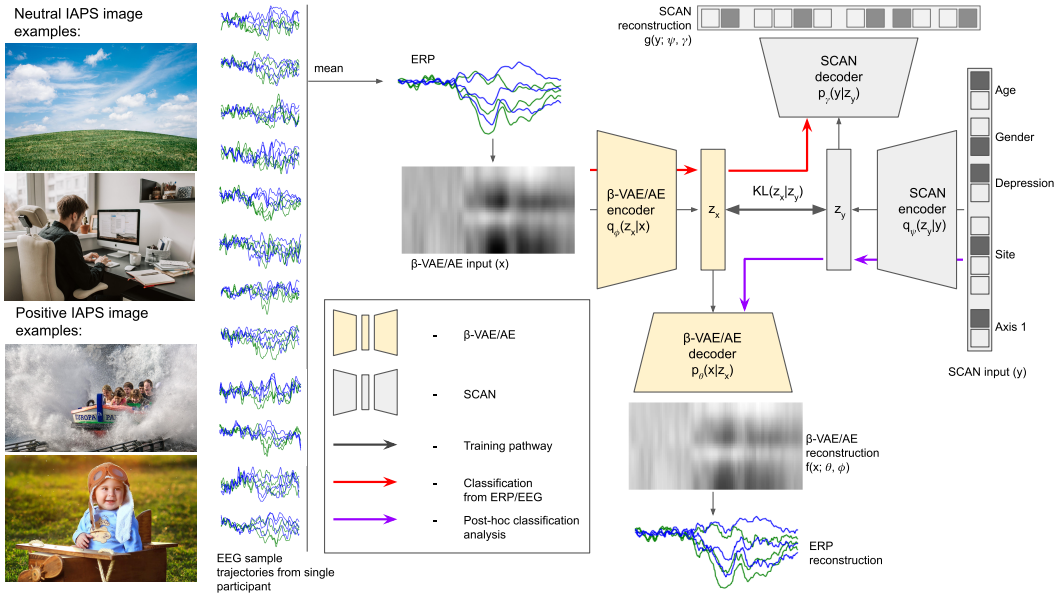

Figure 1: Pipeline schematic. Participants are presented with images from the International Affective Picture System (IAPS). EEG trajectories recorded from the same participant over multiple trials are averaged to create ERPs. Each EEG sample consists of 256 ms of stimulus-locked activity recorded from three sites as participants view either neutral (green) or positive (blue) IAPS images. ERP responses to positive and negative images are concatenated and normalised between [0, 1]. The resulting "images" are used to train β-VAE or AE. A well disentangled pre-trained β-VAE model is used to train SCAN. Each SCAN training example consists of a 5-hot binary classification label $\mathbf{y}$ presented to the SCAN encoder, and its corresponding EEG "image" $\mathbf{x}$ presented to the β-VAE encoder. β-VAE weights are fixed during SCAN training. To obtain classification, an EEG "image" is presented to the β-VAE encoder, then the inferred $\mathbf{z}_x$ means are fed through the SCAN decoder, where a per-class softmax is applied to obtain the predicted label (red pathway). To analyse SCAN classification decisions, a 1-hot binary vector is fed into SCAN encoder. Samples from the inferred distribution $\mathbf{z}_y$ are then fed through the β-VAE decoder to visualise the corresponding ERP reconstructions (purple pathway).

quality is usually a laborious, semi-automated process executed by lab technicians with extensive training. A typical EEG analysis pipeline consists of collecting EEG recordings evoked from a large number of stimulus presentations (trials) in order to have sufficient data to average out the noise. Independent Components Analysis (ICA) is often used to visually identify and remove the component that corresponds to eye blinks (Delorme & Makeig, 2004; Makeig et al., 2004; Jung et al., 2000). This can be followed by a trial rejection stage where anomalous trials are identified and removed from the EEG data scroll, sometimes also through visual examination. The cleaned up EEG recordings from a large number of trials are then averaged to produce an Event Related Potential (ERP) (Luck, 2012). This allows a clinician to extract specific ERP components relevant to the clinical factor of interest, average out the event-locked activity within them, and then either perform a statistical group comparison, or—in the case of the diagnosis classification goal—apply an off-the-shelf classifier, like Logistic Regression (LR) to obtain the final diagnostic results. Some more advanced classification approaches might include Support Vector Machines (SVM), Linear Discriminant Analysis (LDA), or Random Forest (RF) (Parvar et al., 2015; Güler & Übeyli, 2007; Subasi & Gursoy, 2010; Tomioka et al., 2007; Bashivan et al., 2016), but these are currently not common in the field.

To summarise, EEG recordings are noisy measures of electric activity from across the brain. There is evidence that these signals are useful as markers of depression, but we lack understanding of what aspects of depression they index. Furthermore, the field of clinical psychopathology still lacks consensus on the etiopathogenesis of mental health disorders, which means that there is no such thing as the "ground truth" diagnostic labels. Hence, while EEG is routinely used to diagnose conditions like epilepsy (Smith, 2005), memory (Stam et al., 1994) or sleep disorders (Karakis et al., 2012), its promise for being a reliable diagnostic tool for clinical conditions like depression has not been fully realised so far. In order to make EEG a viable diagnostic tool for a broader set of clinical conditions it is important to develop an automated pipeline for extracting the relevant interpretable biomarker correlates from the (preferably individual trial) EEG data in a robust manner. Furthermore, this process should not depend fully on diagnostic labels which are often subjective and at best represent highly heterogeneous classes.

Recent advances in deep learning have prompted research into end-to-end classification of EEG signal using convolutional and/or recurrent neural networks (Bashivan et al., 2016; Mirowski et al., 2009; Cecotti & Graser, 2011; Güler et al., 2005; Wang et al., 2018; Farahat et al., 2019; Solon et al., 2019; Cecotti et al., 2014), holding the promise of automated extraction of the relevant biomarker correlates. However, deep classifiers operate best in the big data regime with clean, well-balanced ground truth classification targets. In contrast, even the largest of EEG datasets typically contain only a few hundred datapoints, and the classification labels are subjective, noisy and unbalanced, with the majority of the data coming from healthy control participants. Hence, in order to utilise the benefits of deep learning but avoid the pitfalls of over-reliance on the classification labels, we propose a two-step pipeline consisting of unsupervised representation learning, followed by supervised mapping of the pre-trained representation to the latest version of the available diagnostic labels. The hope is that the unsupervised learning step would denoise the input signal and extract the broad statistical regularities hidden in it, resulting in a representation that can continue to be useful even if the label taxonomy evolves. To the best of our knowledge, no study yet has successfully applied modern deep representation learning techniques to the problem of identifying clinical biomarkers in the EEG data.

Recently great progress has been made in the field of deep unsupervised representation learning (Devlin et al., 2018; Brown et al., 2020; Chen et al., 2020b; Grill et al., 2020; Chen et al., 2020a; Higgins et al., 2017; Burgess et al., 2019). Disentangled representation learning is a branch of deep unsupervised learning that produces interpretable factorised low-dimensional representations of the training data (Bengio et al., 2013; Higgins et al., 2017). Given the requirement for model interpretability in our use-case, we use Beta Variational Autoencoders (β-VAE) (Higgins et al., 2017)—one of the state of the art unsupervised disentangled representation learning methods—to discover low-dimensional disentangled representations of the EEG data. We then train the recently proposed Symbol–Concept Association Network (SCAN) (Higgins et al., 2018) to map the available classification labels to the representations learnt by β-VAE (see Fig. 1). We demonstrate that our proposed pipeline results in better classification accuracy than the heavily hand-engineered baseline typical for the field. This holds true when predicting a number of factors, including age, gender, and depression diagnosis. Furthermore, SCAN is able to produce highly interpretable classification recommendations, whereby its decisions on different clinical factors are grounded in a small number (often single) latent dimensions of the β-VAE, allowing the clinicians an opportunity to interpret the recommendations produced by SCAN, and visualise what aspects of the EEG signal are associated with the classification decision post-hoc. This opens up the opportunity to use our proposed pipeline as a tool for discovering new EEG biomarkers. We validate this by "re-discovering" a known biomarker for depression. Finally, we demonstrate that once a β-VAE is pre-trained on ERP signals, it can often produce ERP-like reconstructions even when presented with single noisy EEG trajectories. Furthermore, the representations inferred from single EEG trials produce good classification results, still outperforming the canonical baseline method. This suggests that once a good disentangled representation is learnt, the model can be used online as new EEG data is being recorded, thus lowering the burden of keeping potentially vulnerable participants in the lab for extended recording sessions.

2 METHODS

2.1 DATA

Anhedonia. This work targets one of the two cardinal symptoms of depression—anhedonia. Anhedonia is the lack of pleasure and/or interest in previously pleasurable stimuli and activities (DSM-V, 2013). One established approach for objectively quantifying this symptom is the use of EEG to measure neural responses elicited by emotionally salient visual stimuli. Research in this domain has uncovered a stereotyped neural activation pattern in healthy control participants, where emotionally-salient stimuli evoke a Late Positive Potential (LPP) in ERPs—the averaged timeseries of stimulus time-locked EEG recordings. This pattern has been identified as a potential biomarker for depression because (on average) this positive deflection in amplitude is attenuated or absent in individuals who exhibit symptoms of anhedonia (Brush et al., 2018; Foti et al., 2010; Klawohn et al., 2020; MacNamara et al., 2016; Weinberg et al., 2016; Weinberg & Shankman, 2017).

Participants. The data were collected as part of a mental health study across multiple laboratory sites. The multi-site aspect of the study meant that more data could be pooled together, however, it also meant that the data was noisier. Participants ($N = 758$, $\text{age}_{\bar{X}} = 16.7$, $\text{age}_{\text{range}} = [11.0, 59.8]$,

398 female) were samples of healthy controls ($n_{\text{HC}} = 485$) and people diagnosed with anxiety and depression (among other mental illnesses) (see Sec. A.1.1 and Tbl. A1 for further breakdown).

Stimuli and Experimental Design. Participants were shown a series of 80 images from the International Affective Picture System (IAPS) (Lang et al., 2008) presented in random order up to 40 times each. The images varied in valence: either positive (affiliative scenes or cute animals designed to elicit the LPP ERP component), or neutral (objects or scenes with people). Each image trial consisted of a white fixation cross presented for a random duration between 1000-2000 ms (square window) followed by a black and white image presented for 1000 ms.

EEG Preprocessing. While participants completed the picture viewing task, EEG was continuously recorded. Each picture trial was then segmented to contain a 200 ms pre-stimulus baseline and a 1200 ms post-stimulus interval. The raw EEG signal was digitized, bandpass filtered and cleared of the eye movement artifacts and anomalous trials as described in Sec. A.1.2.

Classification labels. The following classification labels were used in this study: age (adult or child), gender (male or female), study site, and the presence or absence of two clinical conditions: depression diagnosis and a broader Axis 1 disorder diagnosis. All classification labels were binary, apart from study site, which contained four possible values corresponding to four different sites where the data were collected. Participants 18 years of age and older were classified as adults. Gender was classified based on self-reported values. Positive depression labels ($n = 110$) include all participants that were diagnosed with Major Depressive Disorder (MDD), Persistent Depressive Disorder (PDD), and depressive disorder NOS (not-otherwise-specified) by expert clinicians through a clinical interview (e.g., SCID for adults, KSADS for children). Axis 1 is a broad category of disorders (now discontinued in DSM-V) that excludes intellectual disabilities and personality disorder (DSM-V, 2013). Positive Axis 1 labels ($n = 273$) encompassed all participants with positive depression labels plus individuals diagnosed with Cyclothemia, Dysthemia, anxiety disorders, mood disorders (e.g., Bipolar Disorder), panic disorders, and substance and eating disorders (all of which are sparse).[1]

2.2 Representation learning

Canonical LPP analysis baseline The canonical approach for extracting the Late Positive Potential (LPP) effect serves as the baseline for this work. The LPP effect is calculated as the average amplitude difference between ERP waveforms (i.e., averaged stimulus time-locked EEG segments) evoked from emotionally-salient and neutral stimuli. Before the delta was calculated, each ERP was normalised with respect to the baseline average activity within the 100 ms window preceding stimulus onset. Finally, the normalised delta signal within the 300–700 ms window after stimulus onset was averaged to provide the canonical baseline LPP representation.

Autoencoder An AutoEncoder (AE) is a deep neural network approach for non-linear dimensionality reduction (Hinton & Salakhutdinov, 2006; Baldi, 2012). A typical architecture consists of an encoder network, which projects the input high-dimensional data $\mathbf{x}$ into a low-dimensional representation $\mathbf{z}$, and a decoder network that is the inverse of the encoder, projecting the representation $\mathbf{z}$ back into the reconstruction of the original high-dimensional data $f(\mathbf{x}; \phi, \theta)$, where ϕ and θ are the parameters of the encoder and decoder respectively (see Fig. 1 for model schematic, and Sec. A.2.1 for model architecture details). AE is trained through backpropagation (Rumelhart et al., 1986) using the reconstruction objective:

$$\mathcal{L}_{AE} = \mathbb{E}_{p(\mathbf{x})} \, ||f(\mathbf{x}; \phi, \theta) - \mathbf{x}||^2$$

The input to the AE in our case is a 256x6 "image" of the EEG signal (see Fig. 1), where 256 corresponds to the 1024 ms of the recorded EEG trajectory sampled at 250 Hz and pre-processed

[1] Disorders included in the Axis 1 label that occur in the data include Major Depressive Disorder, Persistent Depressive Disorder, Depression NOS, Cyclothemia, Dysthemia, Bipolar I, Bipolar II, Bipolar NOS, Mania, Hypomania, Agoraphobia, Social Phobia, Separation Anxiety, Generalized Anxiety Disorder, Panic Disorder, Panic Disorder with Agoraphobia, Anorexia, Bulimia, Eating disorder NOS, Alcohol Abuse, Alcohol Dependence, and Substance Abuse and Substance Dependence disorders.

	Train ↓	Test →	Age		Gender		Site		Depression		Axis 1		Overall	
			ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL	ERP	SMPL
Chance			0.50		0.50		0.25		0.50		0.50		0.45	
LR	ERP	LPP	0.74	0.56	0.52	0.51	0.38	0.30	0.64	0.56	0.52	0.52	0.56	0.49
LR	SMPL	LPP	0.69	0.54	0.52	0.52	0.36	0.30	0.61	0.54	0.52	0.52	0.54	0.48
LR	ERP	β-VAE	**0.85**	0.62	0.66	0.53	**0.51**	0.34	**0.68**	**0.59**	**0.60**	**0.55**	**0.66**	**0.53**
		AE	**0.85**	0.62	**0.68**	**0.57**	0.48	**0.37**	**0.69**	**0.59**	**0.60**	**0.56**	**0.66**	**0.54**
LR	SMPL	β-VAE	0.76	**0.65**	0.58	0.53	0.40	**0.36**	0.67	**0.58**	0.56	**0.56**	0.59	**0.54**
		AE	0.76	**0.65**	0.63	**0.57**	0.46	**0.36**	0.66	**0.58**	**0.59**	0.54	0.62	**0.54**
SCAN	ERP	β-VAE	0.78	**0.64**	0.58	0.50	0.39	0.30	**0.68**	**0.59**	0.54	0.52	0.59	0.51
		AE	0.48	0.50	0.48	0.50	0.21	0.25	0.50	0.49	0.50	0.50	0.43	0.45
SCAN	SMPL	β-VAE	0.73	0.59	0.57	0.53	0.38	0.30	0.67	0.56	0.55	0.50	0.58	0.5
		AE	0.41	0.45	0.43	0.46	0.16	0.20	0.46	0.49	0.53	0.49	0.40	0.42

Table 1: Balanced classification accuracy calculated as the average of precision and recall. ERP - averaged trajectories, SMPL - single sampled EEG trials. LPP - canonical late positive potential baseline representation. LR - L2 regularised logistic regression (see Tbl. A2 for other baseline classifiers). Models trained on single EEG trajectories (SMPL) are tested on different single EEG trajectories in the SMPL/SMPL train/test case.

as described in Sec. A.1.2, and 6 corresponds to three electrodes per each of the two image valence conditions (i.e., stimulus classes): neutral and positive. These input images were further normalised to the [0, 1] range before being presented to the AE.

β-Variational Autoencoder A β-Variational Autoencoder (β-VAE) (Higgins et al., 2017) is a generative model that aims to learn a disentangled latent representation $\mathbf{z}$ of input data $\mathbf{x}$ by augmenting the Variational Autoencoder (VAE) framework (Rezende et al., 2014; Kingma & Welling, 2014) (see Sec. A.2.2 for more details) with an additional β hyperparameter. Intuitively, a disentangled representation is a factorised latent distribution where each factor corresponds to an interpretable transformation of the training data (e.g. in a disentangled representation of a visual scene, individual factors may represent a change in lighting or object position). A neural network implementation of a β-VAE consists of an inference network (equivalent to the AE encoder), that takes inputs $\mathbf{x}$ and parameterises the (disentangled) posterior distribution $q(\mathbf{z}|\mathbf{x})$, and a generative network (equivalent to the AE decoder) that takes a sample from the inferred posterior distribution $\hat{\mathbf{z}} \sim \mathcal{N}(\mu(\mathbf{z}|\mathbf{x}), \sigma(\mathbf{z}|\mathbf{x}))$ and attempts to reconstruct the original image (see Fig. 1). The model is trained through a two-part loss objective:

$$\mathcal{L}_{\beta-VAE} = \mathbb{E}_{p(\mathbf{x})}[\, \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}[\log p_\theta(\mathbf{x}|\mathbf{z})] - \beta KL(q_\phi(\mathbf{z}|\mathbf{x}) \,||\, p(\mathbf{z}))\,]$$

where $p(\mathbf{x})$ is the probability of the input data, $q(\mathbf{z}|\mathbf{x})$ is the learnt posterior over the latent units given the data, $p(\mathbf{z})$ is the unit Gaussian prior with a diagonal covariance matrix $\mathcal{N}(\mathbf{0}, \mathbb{I})$, ϕ and θ are the parameters of the inference (encoder) and generative (decoder) networks respectively, and β is a hyperparameter that controls the degree of disentangling achieved by the model during training. Intuitively, the objective consists of the reconstruction term (which aims to increase the log-likelihood of the observations) and a compression term (which aims to reduce the Kullback-Leibler (KL) divergence between the inferred posterior and the prior). Typically a $\beta > 1$ is necessary to achieve good disentangling, however the exact value differs for different datasets. In order to find a good value of β to disentangle our EEG dataset, we perform a hyperparameter search, training ten models with different random initialisations for each of the ten values of $\beta \in [0.075, 2.0]$ sampled uniformly. Well disentangled β-VAE models were selected using the Unsupervised Disentanglement Ranking (UDR) score (Duan et al., 2019) described in Sec. A.4 (see Fig. A1 for a visualisation of the resulting UDR scores). All β-VAE models had the same architecture as the AE[2], and were trained in the same manner and on the same data.

2.3 Classification

Baseline classifiers To evaluate the quality of a representation in terms of how useful it is for classifying different clinical factors, we applied a range of baseline classification algorithms: Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Linear Discriminant Analysis (LDA) (see Sec. A.3 for details). For all classification results we report balanced accuracy, which is the average of precision and recall calculated according to: $Acc = \left(\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right)/2$. Balanced accuracy was chosen, because it correctly matches chance accuracy even for unbalanced datasets.

[2]Note that the AE can be seen as a special case of the β-VAE with $\beta = 0$.

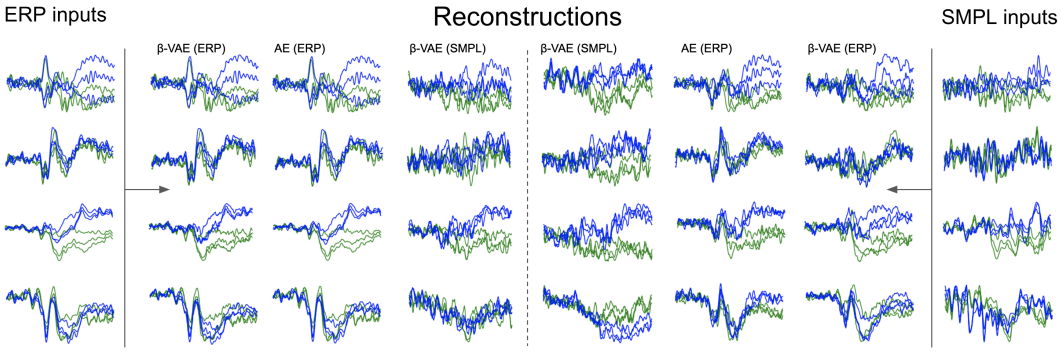

Figure 2: Reconstructions (middle six columns) by two disentangled β-VAEs pre-trained on either ERPs or single EEG sample trajectories (SMPL), as well as an AE pre-trained on ERPs. Each model reconstructs either single EEG trajectories (SMPL inputs) or the corresponding ERPs (ERP inputs). Models pre-trained on ERPs are able to reconstruct ERP-like trajectories even from single EEG samples, as demonstrated by the closer similarity of the β-VAEs (ERP) and AE (ERP) reconstructions to ERP inputs compared to SMPL inputs, regardless of whether ERPs or single EEG samples are reconstructed.

Symbol–Concept Association Network The baseline classifiers described above produced uninterpretable decisions. This is undesirable if we were to have a chance at discovering new clinical biomarkers in the EEG data. To address this interpretability challenge, we leverage a recent model proposed for visual concept learning in the machine learning literature—the Symbol–Concept Association Network (SCAN) (Higgins et al., 2018). While SCAN was not originally developed with the classification goal in mind, it has desirable properties to utilise for the current application. In particular, it is able to automatically discover sparse associative relationships between discrete symbols (5-hot classification labels in our case, see Fig. 1) and continuous probability distributions of the disentangled posterior representation discovered by a trained β-VAE model. Furthermore, the associative nature of the grounding used to train SCAN allows it to deal with noisy data gracefully, and to learn successfully from a small number of positive examples and from highly unbalanced datasets. SCAN is in effect another VAE model. In our case it takes 5-hot classification labels $\mathbf{y}$ as input, and aims to reconstruct them from the inferred posterior $q(\mathbf{z}|\mathbf{y})$ (see Fig. 1 and Sec. A.2.3 for more details). To train SCAN the original VAE objective is augmented with an additional KL term that aims to ground the SCAN posterior in the posterior of a pre-trained β-VAE model:

$$\mathcal{L}_{SCAN} = \mathbb{E}_{p(\mathbf{y})}[\,\mathbb{E}_{q_\psi(\mathbf{z}_y|\mathbf{y})}[\log p_\gamma(\mathbf{y}|\mathbf{z}_y)] - KL(q_\psi(\mathbf{z}_y|\mathbf{y})\,||\,p(\mathbf{z}_y))\,] - KL(q_\phi(\mathbf{z}_x|\mathbf{x})\,||\,q_\psi(\mathbf{z}_y|\mathbf{y}))\,]$$

where ψ and γ are the parameters of the SCAN encoder and decoder respectively, $q_\psi(\mathbf{z}_y|\mathbf{y})$ is the posterior of SCAN inferred from a 5-hot label $\mathbf{y}$, and $q_\phi(\mathbf{z}_x|\mathbf{x})$ is the posterior of the pre-trained β-VAE model inferred from an EEG "image" corresponding to the label presented to SCAN. Note that the β-VAE weights are not updated during SCAN training. The extra grounding term $KL(q_\phi(\mathbf{z}_x|\mathbf{x})\,||\,q_\psi(\mathbf{z}_y|\mathbf{y}))$ allows SCAN to discover an associative relationship between a subset of disentangled factors and each 1-hot dimension of the given label.

3 Results

Deep representation learning improves classification accuracy. We evaluated the classification accuracy for predicting participant age, gender, study site, and the presence or absence of two clinical labels: depression and Axis 1. We first obtained the canonical LPP representations, and applied Support Vector Machine (SVM), Random Forest (RF), Logistic Regression (LR) and Linear Discriminant Analysis (LDA) to obtain balanced classification accuracy. Table A2 shows that LR produced the best results overall for the baseline, hence we only applied LR to the β-VAE and AE representations. Table 1 (ERP/ERP train/test cells) demonstrates that representations extracted through β-VAE and AE pretraining resulted in higher overall classification accuracy than those obtained through the baseline canonical LPP pipeline. This effect holds across all classification tasks, including depression and Axis 1 diagnoses. Furthermore, the pattern of classification results is in line with what might have been expected by the expert clinicians, whereby age and depression classification accuracy is significantly higher than chance, while gender and site are harder to decode from the EEG signal. Indeed, the

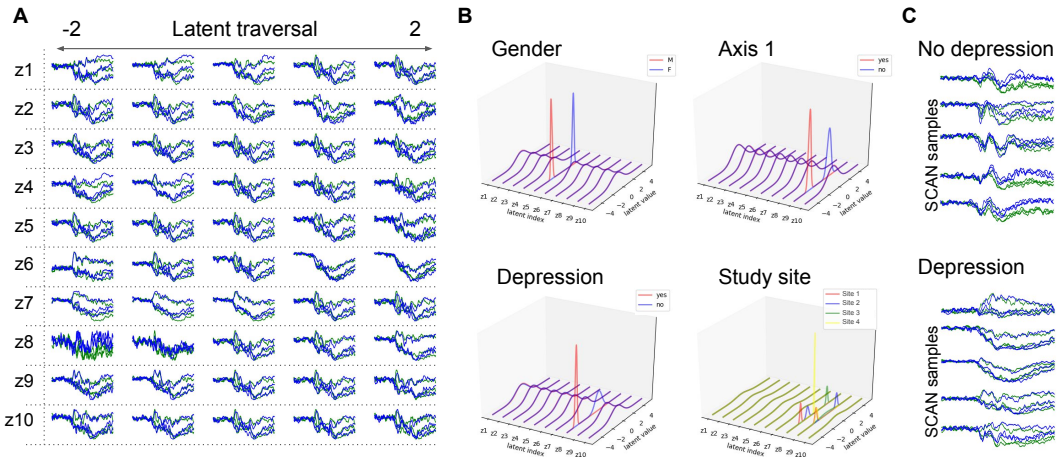

Figure 3: **A**: Latent traversals of a pre-trained disentangled β-VAE. Each row reconstructs an ERP trajectory as the value of each latent dimension is traversed between [-2, 2] while keeping the values of all other latents fixed. **B**: Inferred distributions for each of the 10 latent dimensions of a SCAN trained with the β-VAE shown in A. Each subplot shows the result of SCAN inference on a set of 1-hot symbols corresponding to all the different conditions of a single factor (e.g. M/F for Gender). Different factors have automatically become associated with different (often single) latent dimensions. For example, the presence (red) or absence (blue) of Axis 1 diagnosis is represented by the narrow inferred Gaussian distributions centered around different opposite values of latent z_9, while other latent dimensions are defaulted back to the unit Gaussian prior. **C**: five samples from each of the two conditions associated with the depression factor from the SCAN shown in B. Green - neutral responses, blue - pleasant responses from each of the three EEG sensors. SCAN has automatically discovered that the presence of depression is associated with the lack of clear separation between responses in the two valence conditions—anhedonia.

only way to differentiate between study sites is by picking up on the noise artifacts produced by the differences in the ambient environment and other factors listed in Sec. 1, which are hard to detect.

A similar pattern holds when all the models are trained on single EEG trials instead of ERPs (SMPL/ERP train/test cells in Tbl. 1). While the maximum classification accuracy tends to drop for all models when trained from the noisier single EEG trials rather than ERPs, the classification accuracy obtained from β-VAE and AE representations is still higher than that obtained from the LPP baseline.

Classification based on deep representations generalises better to single EEG trials. One possible clinical application of the proposed classification pipeline is to enable online diagnosis recommendations from single EEG trials. Hence, we tested how well the classification accuracy of the different representations generalises to novel single EEG trial trajectories (SMPL columns in Tbl. 1). As expected, the performance tends to drop on average when classifying the noisy single EEG trajectories, regardless whether the representations were pre-trained using ERPs (ERP rows) or single EEG samples (SMPL rows, note that we tested the models with different single EEG samples to those used for training). However, the classification accuracy is still significantly higher for deep representations compared to the LPP baseline. This suggests that replacing the more manual canonical LPP pipeline with deep representation learning can allow for both better training data efficiency and a reduction in time that the (potentially vulnerable) participants have to spend in the lab by up to 37x, which is the average number of trials per condition that made it through EEG pre-processing in our dataset and were used for generating ERPs.

Deep representation learning reconstructs ERPs from single EEG trials. Since β-VAE and AE had good classification accuracy when presented with single EEG samples, we tested whether they could also reconstruct ERPs from single EEG trials. Figure 2 shows that this is indeed the case—reconstructions produced by the pre-trained models from single noisy EEG samples look remarkably similar to those produced from the corresponding ERPs (also see Figs. A2-A4 for more examples). Note that this only holds true for models trained using ERP data and not those trained on single EEG samples.

Disentangled representations allow for interpretable classification. To obtain SCAN classification results, we inferred the posterior $q_\phi(\mathbf{z}_x|\mathbf{x})$ of a pre-trained β-VAE in response to an EEG "image" $\mathbf{x}$. We then used the pre-trained SCAN decoder to obtain the reconstructed label logits $p_\gamma(\mathbf{y}|\mu(\mathbf{z}_x))$

Age		Gender		Site		Depression		Axis 1	
r	*p*	*r*	*p*	*r*	*p*	*r*	*p*	*r*	*p*
0.28	.01	-0.1	.31	0.19	.06	0.33	<.001	0.31	<.001

Table 2: Spearman correlation between disentanglement quality (as approximated by UDR score) of trained β-VAE models (n=100) and balanced classification accuracy from their corresponding SCAN models.

using the β-VAE posterior mean as input (Fig. 1, red path). Finally, we applied softmax over the produced logits to obtain the predicted 5-hot label for the EEG "image". When SCAN was trained on top of a well disentangled β-VAE, it was able to outperform the canonical LPP baseline in terms of classification accuracy (see Tbl. 2, SCAN+β-VAE rows). This is despite the fact that SCAN was not developed with the classification goal in mind. Note, however, that SCAN can only work when trained with disentangled representations. Indeed, SCAN classification performance was not distinguishable from chance when trained on top of entangled AE representations (see Tbl. 1, SCAN+AE rows). To further confirm the role of disentanglement, we calculated Spearman correlation between the quality of disentanglement as measured by the UDR score and the final SCAN classification accuracy. Table 2 shows significant correlation for age, depression and Axis 1 diagnoses, suggesting that on average these factors were classified better if representations were more disentangled. The same, however, is not the case for gender or study site. This implies that some of the better disentangled models did not contain information necessary for classifying these factors. Such information loss is in fact a known trade-off of β-VAE training, with more disentangled β-VAE models often compromising on the informativeness of the learnt representation due to the increased compression pressure induced by the higher β values necessary to achieve disentanglement (Higgins et al., 2017; Duan et al., 2019).

When trained on top of well disentangled β-VAE models, SCAN decisions were not only accurate, but they were also based on a small number of disentangled dimensions. This is unlike the LR classifier, which obtained average sparsity of just 3.25% compared to the 87.5% for SCAN, even when L1 regularisation was used. Furthermore, the disentangled dimensions used by SCAN were interpretable, hence making SCAN classification decisions amenable to post-hoc analysis. Figure 3B visualises the inferred SCAN posterior when presented with 1-hot labels corresponding to the male or female gender, one of the four study sites, and the presence or absence of depression and Axis 1 diagnoses. In most cases, SCAN was able to associate the label with a single latent dimension in the pre-trained β-VAE (e.g., gender is represented by latent dimension z_5), and the different values of the same class label corresponded to disjoint narrow Gaussians on those latents (e.g., male gender is represented with a Gaussian $\mathcal{N}(0.89, 0.49)$, while female gender is represented by a Gaussian $\mathcal{N}(-0.82, 0.52)$, both on z_5). We can also visualise what the different β-VAE latents have learnt to represent by plotting their traversals as shown in Figure 3A (see more in Figs. A5-A7).

Finally, we can also sample from the SCAN posterior $\hat{\mathbf{z}}_y \sim \mathcal{N}(\mu(\mathbf{z}_y|\mathbf{y}), \sigma(\mathbf{z}_y|\mathbf{y}))$ and reconstruct the resulting ERPs using the β-VAE decoder $p_\theta(\mathbf{x}|\hat{\mathbf{z}}_y)$ (Fig. 1, purple path). Such samples for the depression label are visualised in Figure 3C. It is clear that the reconstructed ERP samples corresponding to participants with a positive depression diagnosis contain no difference between the positive and neutral trials (overlapping blue and green lines), while those corresponding to participants without a positive depression diagnosis do have an obvious gap between the responses to positive and neutral trials. Hence, we were able to "re-discover" an objective biomarker for the symptom of anhedonia (and depression generally), thus opening up the potential for discovering new biomarkers hidden in EEG data in the future.

4 Conclusions

This work provides the first evidence that disentanglement-focused representation learning and the resulting models are powerful measurement tools with immediately applicability to applied and basic research for clinical psychology and electrophysiology. In particular, we have demonstrated that disentangled representation learning can be successfully applied to EEG data, resulting in representations that can be successfully re-used to predict multiple clinical factors through fast supervised re-mapping, outperforming a hand-engineered baseline typical for the field. Furthermore, our method recovers a representation that can be used to automatically extract denoised Event Related Potentials (ERPs) from novel, single EEG trajectories. Finally, single factors of the learned disentangled representations often correspond to meaningful markers of clinical factors (as automatically detected by SCAN), allowing for human interpretability and potentially providing novel insight into new biomarkers and the neurofunctional alterations underlying mental disorders.

REFERENCES

Pierre Baldi. Autoencoders, unsupervised learning, and deep architectures. In *Proceedings of ICML workshop on Unsupervised and transfer learning*, pp. 37–49, 2012.

Pouya Bashivan, Irina Rish, Mohammed Yeasin, and Noel Codella. Learning representations from EEG with deep recurrent-convolutional neural networks. *ICLR*, 2016.

Yoshua Bengio, Grégoire Mesnil, Yann Dauphin, and Salah Rifai. Better mixing via deep representations. In *International Conference on Machine Learning*, pp. 552–560, 2013.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

Christopher J Brush, Peter J Ehmann, Greg Hajcak, Edward A Selby, and Brandon L Alderman. Using multilevel modeling to examine blunted neural responses to reward in major depression. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(12):1032–1039, 2018.

Christopher P Burgess, Loic Matthey, Nicholas Watters, Rishabh Kabra, Irina Higgins, Matt Botvinick, and Alexander Lerchner. MoNET: Unsupervised scene decomposition and representation. *arXiv preprint arXiv:1901.11390*, 2019.

Hubert Cecotti and Axel Graser. Convolutional neural networks for P300 detection with application to brain-computer interfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):433–445, 2011.

Hubert Cecotti, Miguel P Eckstein, and Barry Giesbrecht. Single-trial classification of event-related potentials in rapid serial visual presentation tasks using supervised spatial filtering. *IEEE transactions on Neural Networks and Learning Systems*, 25(11):2030–2042, 2014.

Mark Chen, Alec Radford, Rewon Child, Jeff Wu, Heewoo Jun, Prafulla Dhariwal, David Luan, and Ilya Sutskever. Generative pretraining from pixels. In *Proceedings of the 37th International Conference on Machine Learning*, 2020a.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. A simple framework for contrastive learning of visual representations. *arXiv preprint arXiv:2002.05709*, 2020b.

Arnaud Delorme and Scott Makeig. EEGLAB: an open source toolbox for analysis of single-trial EEG dynamics including independent component analysis. *Journal of Neuroscience Methods*, 134 (1):9–21, 2004.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

DSM-V. *Diagnostic and statistical manual of mental disorders*. American Psychiatric Association, Washington, DC, 5 edition, 2013.

Sunny Duan, Loic Matthey, Andre Saraiva, Nicholas Watters, Christopher P Burgess, Alexander Lerchner, and Irina Higgins. Unsupervised model selection for variational disentangled representation learning. *arXiv preprint arXiv:1905.12614*, 2019.

Amr Farahat, Christoph Reichert, Catherine M Sweeney-Reed, and Hermann Hinrichs. Convolutional neural networks for decoding of covert attention focus and saliency maps for EEG feature visualization. *Journal of Neural Engineering*, 16(6):066010, 2019.

Dan Foti, Doreen M Olvet, Daniel N Klein, and Greg Hajcak. Reduced electrocortical response to threatening faces in major depressive disorder. *Depression and Anxiety*, 27(9):813–820, 2010.

Eiko I Fried and Randolph M Nesse. Depression is not a consistent syndrome: an investigation of unique symptom patterns in the STAR* D study. *Journal of Affective Disorders*, 172:96–102, 2015.

Gabriele Gratton, Michael GH Coles, and Emanuel Donchin. A new method for off-line removal of ocular artifact. *Electroencephalography and Clinical Neurophysiology*, 55(4):468–484, 1983.

Jean-Bastien Grill, Florian Strub, Florent Altché, Corentin Tallec, Pierre H Richemond, Elena Buchatskaya, Carl Doersch, Bernardo Avila Pires, Zhaohan Daniel Guo, Mohammad Gheshlaghi Azar, et al. Bootstrap your own latent: A new approach to self-supervised learning. *arXiv preprint arXiv:2006.07733*, 2020.

Inan Güler and Elif Derya Übeyli. Multiclass support vector machines for EEG-signals classification. *IEEE transactions on Information Technology in Biomedicine*, 11(2):117–126, 2007.

Nihal Fatma Güler, Elif Derya Übeyli, and Inan Güler. Recurrent neural networks employing Lyapunov exponents for EEG signals classification. *Expert Systems with Applications*, 29(3):506–514, 2005.

Greg Hajcak, Julia Klawohn, and Alexandria Meyer. The utility of event-related potentials in clinical psychology. *Annual Review of Clinical Psychology*, 2019.

Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. β-VAE: Learning basic visual concepts with a constrained variational framework. *ICLR*, 2(5):6, 2017.

Irina Higgins, Nicolas Sonnerat, Loic Matthey, Arka Pal, Christopher P Burgess, Matko Bosnjak, Murray Shanahan, Matthew Botvinick, Demis Hassabis, and Alexander Lerchner. SCAN: Learning hierarchical compositional visual concepts. *ICLR*, 2018.

Geoffrey E Hinton and Ruslan R Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

Thomas Insel, Bruce Cuthbert, Marjorie Garvey, Robert Heinssen, Daniel S Pine, Kevin Quinn, Charles Sanislow, and Philip Wang. Research domain criteria (RDoC): toward a new classification framework for research on mental disorders, 2010.

Tzyy-Ping Jung, Scott Makeig, Colin Humphries, Te-Won Lee, Martin J Mckeown, Vicente Iragui, and Terrence J Sejnowski. Removing electroencephalographic artifacts by blind source separation. *Psychophysiology*, 37(2):163–178, 2000.

Ioannis Karakis, Keith H Chiappa, Marta San Luciano, Kenneth C Sassower, John W Stakes, and Andrew J Cole. The utility of routine EEG in the diagnosis of sleep disordered breathing. *J Clin Neurophysiol*, 29(4):333–338, 2012.

Ronald C Kessler, Katherine A McGonagle, Marvin Swartz, Dan G Blazer, and Christopher B Nelson. Sex and depression in the national comorbidity survey i: Lifetime prevalence, chronicity and recurrence. 1993.

Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*, 2014.

Julia Klawohn, Kreshnik Burani, Alec Bruchnak, Nicholas Santopetro, and Greg Hajcak. Reduced neural response to reward and pleasant pictures independently relate to depression. *Psychological Medicine*, pp. 1–9, 2020.

Peter J Lang, Margaret M Bradley, and Bruce N Cuthbert. International affective picture system (IAPS): Technical manual and affective ratings. *The Center for Research in Psychophysiology, University of Florida.*, 1:39–58, 2008.

Grace Y Lim, Wilson W Tam, Yanxia Lu, Cyrus S Ho, Melvyn W Zhang, and Roger C Ho. Prevalence of depression in the community from 30 countries between 1994 and 2014. *Scientific Reports*, 8 (1):1–10, 2018.

Jill Lobbestael, Maartje Leurgans, and Arnoud Arntz. Inter-rater reliability of the Structured Clinical Interview for DSM-IV Axis I disorders (SCID I) and Axis II disorders (SCID II). *Clinical psychology & psychotherapy*, 18(1):75–79, 2011.

Rafael Lozano, Mohsen Naghavi, Kyle Foreman, Mohammad A AlMazroa, and Ziad A Memish. Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010 (vol 380, pg 2095, 2012). *The Lancet*, 381:628–628, 2013.

Steven J Luck. Event-related potentials. 2012.

Annmarie MacNamara, Roman Kotov, and Greg Hajcak. Diagnostic and symptom-based predictors of emotional processing in generalized anxiety disorder and major depressive disorder: An event-related potential study. *Cognitive Therapy and Research*, 40(3):275–289, 2016.

Scott Makeig, Stefan Debener, Julie Onton, and Arnaud Delorme. Mining event-related brain dynamics. *Trends in Cognitive Sciences*, 8(5):204–210, 2004.

Sally McManus, Howard Meltzer, Traolah Brugha, Paul E Bebbington, and Rachel Jenkins. Adult Psychiatric Morbidity in England 2007: results of a household survey. NHS Information Centre for Health and Social Care. *NHS Information Centre for Health and Social Care*, 2009.

Sally McManus, Paul E Bebbington, Rachel Jenkins, and Traolah Brugha. Mental health and wellbeing in England: Adult Psychiatric Morbidity Survey 2014. *Leeds: NHS Digital*, 2016.

Piotr Mirowski, Deepak Madhavan, Yann LeCun, and Ruben Kuzniecky. Classification of patterns of EEG synchronization for seizure prediction. *Clinical Neurophysiology*, 120(11):1927–1940, 2009.

Hossein Parvar, Lauren Sculthorpe-Petley, Jason Satel, Rober Boshra, Ryan CN D'Arcy, and Thomas P Trappenberg. Detection of event-related potentials in individual subjects using support vector machines. *Brain Informatics*, 2(1):1–12, 2015.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*, 2014.

David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

SJM Smith. EEG in the diagnosis, classification, and management of patients with epilepsy. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(suppl 2):ii2–ii7, 2005.

Amelia J Solon, Vernon J Lawhern, Jonathan O Touryan, Jonathan R McDaniel, Anthony J Ries, and Stephen M Gordon. Decoding P300 variability using convolutional neural networks. *Frontiers in Human Neuroscience*, 13:201, 2019.

Kees J Stam, Dénes LJ Tavy, Brechtje Jelles, Herbert AM Achtereekte, Joris PJ Slaets, and Ruud WM Keunen. Non-linear dynamical analysis of multichannel EEG: clinical applications in dementia and Parkinson's disease. *Brain Topography*, 7(2):141–150, 1994.

Abdulhamit Subasi and M Ismail Gursoy. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications*, 37(12):8659–8666, 2010.

Ryota Tomioka, Kazuyuki Aihara, and Klaus-Robert Müller. Logistic regression for single trial EEG classification. In *NeurIPS*, pp. 1377–1384, 2007.

Theo Vos, Ryan M Barber, Jashua A Salomon, and Christopher J L Murray. Global, regional, and national incidence, prevalence, and years lived with disability for 301 acute and chronic diseases and injuries in 188 countries, 1990–2013: a systematic analysis for the Global Burden of Disease Study. *The Lancet*, 386(9995):743–800, 2013.

Haofei Wang, Bertram E. Shi, and Yiwen Wang. Convolutional neural network for target face detection using single-trial EEG signal. *2018 40th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pp. 2008–2011, 2018.

Anna Weinberg and Stewart A Shankman. Blunted reward processing in remitted melancholic depression. *Clinical Psychological Science*, 5(1):14–25, 2017.

Anna Weinberg, Greg Perlman, Roman Kotov, and Greg Hajcak. Depression and reduced neural response to emotional images: Distinction from anxiety, and importance of symptom dimensions and age of onset. *Journal of abnormal psychology*, 125(1):26, 2016.

Harvey A Whiteford, Louisa Degenhardt, Christopher J L Murray, and Theo Vos. Global burden of disease attributable to mental and substance use disorders: findings from the Global Burden of Disease Study 2010. *The Lancet*, 382(9904):1575–1586, 2013.