CHAPTER

15 Randomized Deep Learning Methods for Clinical Trial Enrichment and Design in Alzheimer's Disease

Vamsi K. Ithapu[*], **Vikas Singh**[*], **Sterling C. Johnson**[†,*]

University of Wisconsin–Madison, Madison, WI, United States[*] *William S. Middleton Memorial Hospital, Madison, WI, United States*[†]

CHAPTER OUTLINE

Deep Learning for Medical Image Analysis. DOI: 10.1016/B978-0-12-810408-8.00020-1

15.1 INTRODUCTION

Alzheimer's disease (AD) affects over 20 million people worldwide [1], and in the last decade, efforts to identify AD biomarkers have intensified. There is now broad consensus that the disease pathology manifests in the brain images years before the onset of AD. Various groups have adapted sophisticated machine learning methods, to *learn* patterns of pathology by classifying healthy controls from AD subjects. The success of these methods (which obtain over 90% accuracy [2]) has led to attempts at more fine grained classification tasks, such as separating controls from Mild Cognitively Impaired (MCI) subjects and even identifying which MCI subjects will go on to develop AD [3,4]. Even in this difficult setting, multiple current methods have reported over 75% accuracy. While accurate classifiers are certainly desirable, one may ask if they address a real practical need – if no treatments for AD are currently available, is AD diagnosis meaningful? To this end, [5,6] showed the utility of computational methods beyond diagnosis/prognosis; they may, in fact, be leveraged for designing *efficient* clinical trials for AD. There are, however, several issues with such procedures, and this work is in this context of designing methods for AD trials which are *deployable in practice* and *cost-effective*.

Recent clinical trials designed to evaluate new treatments and interventions for AD at the mild to moderate dementia stage have largely been unsuccessful and there is growing consensus that trials should focus on the earlier stages of AD including MCI or even the presymptomatic stage [7,8], if such stages can be accurately identified in individual subjects [9–11]. However, MCI is a clinical syndrome with heterogeneous underlying etymology that may not be readily apparent from a clinical work-up, posing a major challenge in reliably identifying the most probable beneficiaries of a putative effective treatment [12]. For instance, MCIs may have clinical but not biomarker evidence of incipient AD, may have biomarker evidence in some modalities, or, despite biomarker presence, may not show symptomatic progression during the trial time-period. An efficient MCI trial would ideally include only those patients most likely to benefit from treatment; who possess AD pathology based on a constellation of amyloid, tau and neural injury biomarker assessments, and who are most likely to progress clinically to symptomatic AD. The typical annual conversion rate to dementia among MCI due to AD is 3–20% across several studies [13]. Hence, over a two year trial, at best only 40% of participants would have naturally progressed, and the ability to detect the true efficacy of the trial is perhaps diminished.

Several ongoing AD trials "enrich" their population using one or more disease markers as inclusion criteria [8,14]. The general framework here is to effectively screen out subjects who are weak decliners (i.e., MCI who may not convert to AD) [15]. Unless there is a natural phase change (i.e., an elbow) in the distribution for distinguishing the at-risk and not-at-risk subjects, a fixed fraction of the total cohort are filtered out based on the study design. Imaging-based markers (e.g., fluorodeoxyglucose (FDG), hippocampal and ventricular volume) and cerebrospinal fluid (CSF) profiles have been shown to be effective in screening out low-risk subjects, due to the fact that disease manifests much earlier in imaging data compared to cognition [7,8].

However, these markers are uni-modal while several studies have shown the efficacy of multi-modal data [4,2]. Furthermore, CSF cannot be used in practice as a screening instrument because assays typically need to be performed in a single batch and are highly lab specific [16]. Several recent studies have used *computational* multi-modal markers derived from support vector machines (SVMs) and other machine learning models [14,5,6,17,18]. The strategy here uses imaging data from two time points (i.e., TBM or hippocampus volume change) and derives a machine learning based biomarker. Based on this marker, say, the top (strongest decliners) one-third quantile subjects may be selected to be included in the trial. Using this enriched cohort, the drug effect can then be detected with higher statistical power, making the trial more cost effective and far easier to setup/conduct. Most such approaches use longitudinal data, however, a practical enrichment criterion should only use baseline (trial start-point) data. We argue that existing approaches to enrichment, including state-of-the-art computational techniques, *cannot* guarantee this optimal enrichment behavior – optimally correlate with dementia spectrum with high confidence having access only to the baseline data, while simultaneously ensuring small intra-stage variance.

We approach the optimal enrichment design from basic principles. Specifically, consider a trial where participants are randomly assigned to treatment (intervened) and placebo (non-intervened) groups, and the goal is to quantify any drug effect. Traditionally, this effect is quantified based on a "primary" outcome, like cognitive measure or brain atrophy. If the distributions of this outcome for the two groups are statistically different, we conclude that the drug is effective. When the effects are subtle, the number of subjects required to see statistically meaningful differences can be huge, making the trial infeasible. Instead, one may derive a "customized outcome" from a statistical machine learning model that assigns predictions based on probabilities of class membership (no enrichment is used). If these customized predictions are statistically separated (classification is a special case), it directly implies that potential improvements in power and the efficiency of the trial are possible. This paper is focused on designing specialized learning architectures toward this final objective. In principle, *any* machine learning method should be appropriate for the above task. But it turns out that high statistical power in these experiments is not merely a function of the classification accuracy of the model, rather the conditional entropy of the outputs (prediction variables) from the classifier at test time. An increase in classifier accuracy does not directly reduce the variance in the predictor (from the learned estimator). Therefore, SVM type methods *are* applicable, but significant improvements are possible by deriving a learning model with the *concurrent* goals of classifying the stages of dementia *as well as* ensuring small conditional entropy of the outcomes.

We achieve the above goals by proposing a novel machine learning model based on ideas from deep learning [19]. Deep architectures are nonparametric learning models [20–22] that have received much interest in machine learning, computer vision and natural language processing recently [23–25]. They are effective and robust in learning complex concepts, and recently several authors extensively studied their success from both empirical and theoretical view-points [26,27]. Although powerful, it

is well known that they require very large amounts of data (unsupervised or labeled) [20], which is infeasible in medical applications including neuroimaging, bioinformatics, etc., where data dimensionality (d) is always much larger than the number of instances (n). A naïve use of off-the-shelf deep networks expectedly yields poor performance. Nevertheless, independent of our work, deep learning was used in structural and functional neuroimaging, where [28] use a region of interest approach while [29,30] sub-samples each data instance to increase n. Our work provides a mechanism where no such adjustments are necessary, and, in fact, the framework developed here is more generally applicable for learning deep networks in small sample regime (i.e., learning problems where number of data instances is much smaller than the data dimensionality like whole-brain voxel-wise analysis).

Our contributions. (a) We propose a novel, scalable, and a general, deep learning framework that is applicable for learning problems in the small sample regime, and provide certain guarantees for their performance; (b) Using our proposed models, we design novel predictive multi-modal imaging-based disease markers, based only on the trial start-time (baseline) acquisitions, that correlates very strongly with future AD decline; and (c) We show via extensive analyses using imaging, cognitive and other clinical data that the new computational markers result in cost-efficient clinical trials with moderate sample sizes when used as trial inclusion criteria. Section 15.2 briefly introduces clinical trials and deep networks, and Section 15.3 presents the optimal enrichment design. Section 15.4 presents our proposed models, referred to as *randomized deep networks*. Sections 15.5 and 15.6 extensively evaluate these models and discuss their efficacy in enrichment.

15.2 BACKGROUND

The design and learning of randomized deep networks is directly motivated by the optimal enrichment design. Hence, we first present some background on the sample size estimation for conducting clinical trials [31] discussing the necessity of a *computational enrichment criterion*. Although there are many classes/types of deep architectures in the literature, we focus our presentation using two of the most widely used (and well studied) architectures – stacked denoising autoencoders [32] and fully-supervised dropout [33]. The ideas presented here are applicable for any such architectures used for learning problems in the small-sample regime.

15.2.1 CLINICAL TRIALS AND SAMPLE ENRICHMENT

Consider a randomized clinical trial (RCT) designed to test the efficacy of some treatment for an underlying disease condition [31,34]. The population under study are randomly assigned to either of the treatment (or intervention) and non-treatment (or placebo) groups. If the drug indeed offers an improvement, then the two groups should show this change when measured using some outcome measure summarizing

the disease status. Such change would generally correspond to *reducing* the disease progression to a certain extent, referred to as the effect size [34], in the treatment group compared to the placebos. The outcome measure is, in general, a reliable disease marker. Given such an outcome, the trial efficacy is measured by estimating the Type-II error between the two groups, after inducing the drug or intervention. Clearly, this Type-II error (and the effect size) would be influenced by the choice of the outcome and the size (and demographics) of the trial population. Hence, in practice, one would want to "estimate" the trial's efficacy ahead of time to ensure that the effect size is good enough, the population is reasonably large and diverse, and the outcome is appropriately chosen – basically ensuring that the trial makes sense. In such a hypothetical RCT, the drug is induced by *fixing* the effect size ahead of time, and computing the resulting Type-II error for the given outcome and population.

Let δ denote the difference of mean outcome (the standard change) between the trial start and end points (e.g., 2 years) in the placebos. Let σ be the standard deviation of the outcome, and the effect size be η (e.g., 0.25, a 25% reduction in the disease is desired). δ and σ are known a priori (reported in alternate studies on the disease). The treatment group is then *expected* to have the change in outcome decreased to $(1-\eta)\delta$, which will correspond to a hypothetical improvement of η induced by the drug/treatment. Within this setting, the number of samples s required per arm (treatment and placebo) is given by [31]

$$s = \frac{2(Z_\alpha - Z_{1-\beta})^2\sigma^2}{(1-\eta)^2\delta^2} \tag{15.1}$$

where $(1-\beta)$ denotes the desired statistical power at a significance level of α. This expression directly follows from applying a difference of means t-test, where the means are computed from the two distributions of interest – outcome change in treatment and placebo groups [31]. The null hypothesis is that mean change in the outcome is same for the treatment and placebo groups. The necessity of sample enrichment can be directly seen from (15.1). If the population under study has less standard change in the outcome δ, then the required sample s for achieving a given power $1-\beta$ will be very large. Alternatively, the power can be maximized only when the incipient change δ is as large as possible in the trial population. Hence, the participants need to be *enriched* so as to include only those subjects with large change in the disease *during* the trial time-line. We define an *optimal enrichment criterion*, in Section 15.3, based on these ideas and (15.1).

15.2.2 NEURAL NETWORKS

Artificial neural networks (ANN) are representation learning machines introduced in the late 1950s for learning complex concepts [35]. An ANN transforms a given (input) instance/example (a d-dimensional vector of features/covariates) into a new representation that may be used within the context of classification or regression or other learning paradigms. These transformations are nonlinear, and possibly non-convex,

and calculated by first computing an affine projection of the input, followed by applying a monotonic nonlinear "activation" function over these projections (which may not be necessarily point-wise) [20]. The resulting features correspond to some high-level and abstract representations of the inputs. Given a set of examples $(\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ from n different instances (or subjects in the case of medical imaging), the new representations are given by $\mathbf{h}_i = \sigma(\mathbf{W}\mathbf{x}_i + b)$. $\mathbf{W} \in \mathbb{R}^{d_1 \times d}$ and $b \in \mathbb{R}^{d_1 \times 1}$ are the unknown transformation coefficients (i.e., weights) (assuming $y_i \in \mathbb{R}^{d_1}$). $\sigma(\cdot)$ is the point-wise activation, and in general, it is a sigmoid (i.e., $\sigma(z) = \frac{1}{1+\exp(-z)}$), although other types of functions including hyperbolics or rectified linear units [36] may be used.

This *single-layer* neural network comprises of a visible layer ($\mathbf{x}_i$s) and an output layer (y_is). *Multi-layer* neural networks (MLNN) perform $L > 1$ such transformations sequentially, thereby resulting in L *hidden* layers ($\mathbf{h}_i^l$s, $l = 1, \ldots, L$):

$$\mathbf{h}_i^l = \sigma(\mathbf{W}^l \mathbf{h}_i^{l-1} + b^l), \quad \mathbf{h}_i^0 = \mathbf{x}_i, \quad l = 1, \ldots, L, \quad i = 1, \ldots, n. \tag{15.2}$$

The highest level hidden layer ($\mathbf{h}_i^L$s) may then be used for a given learning task. Specifically, a classification or regression model can be trained using $(\mathbf{h}_i^L, y_i)_{i=1}^n$, or in certain situations, $\mathbf{h}_i^L$s may directly correspond to y_is. Once the learning is done, the output layer of MLNN would simply be the predictions $\hat{y}_i$s. The hidden layer lengths are denoted by $d_l, l = 1, \ldots, L$, and whenever $d_l > d \ \forall l$, the network learns over-representations of the inputs that are invariant and abstract in some sense. Fig. 15.1A shows the architecture of a typical L-layered MLNN. Layer-to-layer connections shown in the figure correspond to applying the corresponding transformations $((\mathbf{W}^l, b^l))$, followed by point-wise sigmoid nonlinearity $\sigma(\cdot)$.

15.2.3 BACKPROPAGATION AND DEEP LEARNING

To learn the unknown transformations $(\mathbf{W}^l, b^l)$ one compares the MLNN predictions $\mathbf{y}_i$s to the desired outputs using some appropriate loss function $\ell(\cdot)$ (e.g., squared loss, entropy, or divergences), chosen entirely based on the problem at hand. This objective is non-convex because of the presence of multiple (nested) composition of nonlinearities, and hence, the estimation proceeds via stochastic gradients on the loss function objective [37]. This procedure is referred to as backpropagation [38], because the errors in the final/output layer need to be 'propagated' back to the inputs for estimating the coefficients. Several variants of backpropagation have been proposed over the last few decades. The reader should note that it is impractical to review (or even refer to) the vast literature on exhaustive empirical analysis of the various backpropagation strategies – and there have been very many. We only point out the main bottlenecks of backpropagation, which eventually made the neural network learning obsolete (while kernel machines and decision trees received more attention because of their ease of learning and implementability). An interested reader can refer to [19,39] and others for further details.

Although MLNNs have attractive theoretical properties (e.g., a 3-layer network can represent any polynomial of arbitrarily high degree, see Chapter 5 in [39]), training them involves a difficult optimization task due to the composition of nonlinear (and possibly non-convex) functions like sigmoids. The parameter space is highly non-convex with many local minima and saddle points [37] (see Chapter 6 in [19]). Stochastic minimization methods, nevertheless, compute a local minima, but these may be sensitive to initialization [20,38]. The *goodness* of such solutions – in terms of stability to noise, saddle point behavior and sensitivity to perturbations – depend on the number of stochastic iterations, and in turn, on the number of training samples available to exhaustively (and empirically) search through the solution space. Further, gradient based methods are prone to exponential decay of errors for large MLNNs, since the errors computed at the last layer "die-out", in some sense, by the time they reaches the inputs because of the presence of nonlinearities. Hence, even large errors from the objective may not lead to reasonably good gradient paths for the bottom layer transformations. Lastly, MLNN learning involves critical modeling/design choices about the network structure (number and lengths of layers). Although several studies have shown the necessity of over-represented networks, there was no consensus on which variant of classical backpropagation would best suite a given choice of network, and if there was any standardized way of choosing the network structure. These issues make efficient and/or robust learning of MLNNs difficult, which will eventually lead to poor generalization. Deep learning refers to a suite of algorithms for efficient training of MLNNs by mitigating some of these issues [20, 19]. The *depth* simply refers to the many levels of transformations.

Recall the main issue with MLNNs was that the multiple function compositions makes the search space non-convex with many local minima and saddle points. Deep learning algorithms partly mitigate this issue by disentangling the compositions and only working with one-layer (i.e., one transformation) at a time. Working with each layer individually makes the objective simpler albeit, non-convex but likely with fewer local minima. Once *good* estimates of the transformations are obtained *layer-wise*, the entire network can then be initialized with these estimates. The resulting final layer predictions can then be compared with desired outputs to adjust or *fine-tune* the estimated parameters across all layers. This fine tuning is the same as performing complete backpropagation but using layer-wise estimates as an initialization or warm-start [20]. The basic rationale is that once reasonable layer-wise warm-starts are obtained, the transformation coefficients are already in some good solution bowl in the gradient search space, which may be smoother and better behaved than the ones obtained with random initializations (and/or learning all layers concurrently) or whole-network warm starts [20]. The challenge then is to construct such efficient layer-wise procedures to perform this two stage learning – initialize with layer-wise *pretraining*, followed by global fine-tuning. Several such procedures have been constructed, all of which broadly fall under the two categories – restricted Boltzmann machines [40] and autoencoders [32]. More recently, an interesting learning procedure referred to as "Dropout", has been proposed to address the over-fitting problem whenever large MLNNs are to be learned [33]. Specifically, [41,24] showed

that one can perform fully supervised dropout with dropout rate of 0.5 (we will discuss more about this shortly) and completely *by-pass* the layer-wise pretraining. As discussed earlier, we present our models using one of the autoencoder schemes, denoising autoencoders, and dropout learning as building blocks.

15.2.3.1 Denoising Autoencoders (DA) and Stacked DA (SDA)

An autoencoder is a single-layer network that learns robust *distributed representations* of the input data. Given inputs $\mathbf{x}_i$, the autoencoder learns hidden/latent representations $\mathbf{h}_i = \sigma(\mathbf{W}\mathbf{x}_i + b)$, such that the reconstructions $\hat{\mathbf{x}}_i = \sigma(\mathbf{W}^T\mathbf{h}_i + c)$ are as *close* as possible to $\mathbf{x}_i$. It minimizes the following input reconstruction error

$$\mathcal{Z}_a(\{\mathbf{x}_i\}_1^n, \theta) := \arg\min_{\mathbf{W},b,c} \sum_{i=1}^{N} \ell(\mathbf{x}_i, \sigma(\mathbf{W}^T\boldsymbol{\sigma}(\mathbf{W}\mathbf{x_i} + \mathbf{b}) + c)) \tag{15.3}$$

where $\ell(\cdot,\cdot)$ denotes a suitable loss function, e.g., squared loss. A stochastic gradient scheme [42] can be used to perform this minimization. Observe that with no other constraints on $(\mathbf{W}, b, c)$, the above minimization could potentially learn *identity* mappings, i.e., $\mathbf{h}_i$s will be identical to the inputs, making the autoencoding setup useless. Several approaches have been suggested to avoid such identity mappings and instead learn useful representations [20]. We consider the approach where the inputs $\mathbf{x}_i$ are *corrupted* stochastically and the autoencoder is forced to reconstruct the original (non-corrupted) versions. This is referred to as a *denoising* autoencoder (DA) [32], and the minimization in (15.3) will change to

$$\mathcal{Z}_{da}(\{\mathbf{x}_i\}_1^n, \theta) := \arg\min_{\mathbf{W},b,c} \sum_{i=1}^{n} \mathbb{E}_{\tilde{\mathbf{x}}\sim\gamma(\tilde{\mathbf{x}}|\mathbf{x})}\ell(\mathbf{x}_i, \sigma(\mathbf{W}^T\sigma(\mathbf{W}\tilde{\mathbf{x}}_i + b) + c)) \tag{15.4}$$

where $\gamma(\cdot)$ is a stochastic corruption function, and $\tilde{\mathbf{x}}_i$ represents the corrupted $\mathbf{x}_i$. $\gamma(x_{ij}) = x_{ij}$ with some (given) probability ζ and 0 elsewhere ($j = 1, \ldots, d$ are the data dimensions).

DA is a stochastic autoencoder whose learning procedure seeks to undo the input corruptions (hence the name, denoising). The corruption forces the transformations to correspond to some properties of input data, since the reconstruction error in (15.4) decreases only if the transformations pick out the most informative data dimensions. Hence $\mathbf{h}_i$s are abstract enough to *generate* the inputs [32]. Fig. 15.1B summarizes the DA learning. Multiple DAs can then be concatenated to construct a *stacked* DA (SDA), where the hidden representations of lth DA are the uncorrupted inputs to $(l+1)$th DA ($l = 1, \ldots, L-1$). The objective of SDA is

$$\begin{aligned} &\mathcal{Z}_{sda}(\{\mathbf{x}_i\}_1^n, L, \theta) := \sum_{l=0}^{L-1} \mathcal{Z}_{da}(\{\mathbf{h}_i^l\}_1^n, \theta); \\ &\mathbf{h}_i^l = \sigma(\mathbf{W}^l\mathbf{h}_i^{l-1} + p^l); \quad \mathbf{h}_i^0 = \mathbf{x}_i. \end{aligned} \tag{15.5}$$

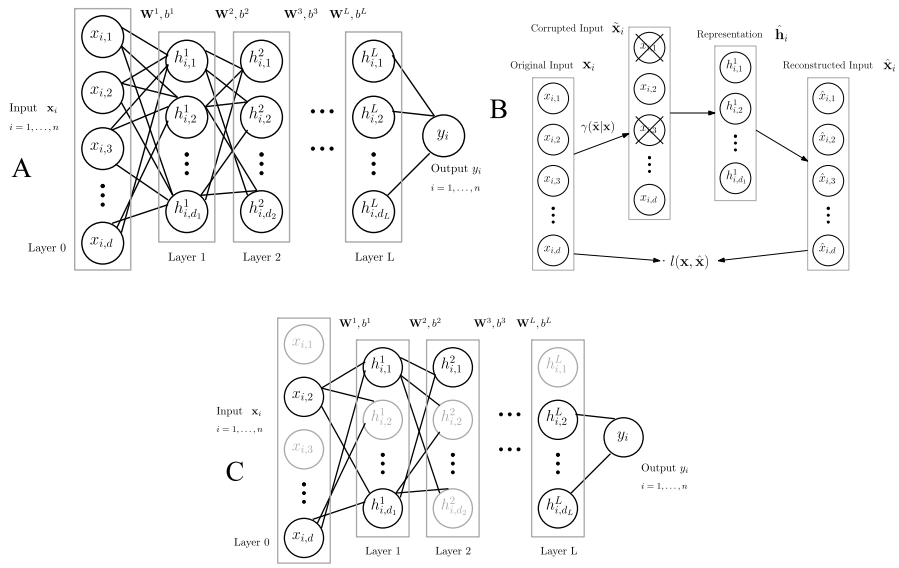

FIGURE 15.1

(A) L-layered MLNN transforming the input features $\mathbf{x}_i$ to a desired output $\mathbf{y}_i$. The lengths of hidden layers are $d_1, d_2, \ldots, d_L$. Layer 0 is the visible layer corresponding to the inputs $\mathbf{x}_i$. Layers 1 to L are the L hidden layers, and y_is denote the outputs. Layer-to-layer transformations are represented by $\mathbf{W}^l$ and b^l. $i = 1, \ldots, n$ and $l = 1, \ldots, L$. (B) The learning process of DA where the input $\mathbf{x}_i$ is corrupted to generate $\tilde{\mathbf{x}}_i$, which is then used to reconstruct an approximation of $\mathbf{x}_i$ denoted by $\hat{\mathbf{x}}_i$. The crossed-out elements in $\tilde{\mathbf{x}}_i$ represented the stochastically corrupted units i.e., for these units $\tilde{\mathbf{x}}_{i,.} = 0$. For the rest of the non-crossed-out units $\tilde{\mathbf{x}}_{i,.} = \mathbf{x}_{i,.}$. The loss function $\ell(\cdot, \cdot)$ then compares the original $\mathbf{x}_i$ to the reconstruction $\hat{\mathbf{x}}_i$. (C) The strategy of dropout learning, where, a fraction $\eta(\cdot)$ of all the units are *dropped* in each layer i.e., all the connections involving these units are dropped (biases are never dropped). Under the given iteration, this smaller network is learned instead of the entire one, and the process repeats stochastically across all iterations.

It is straightforward to see that the structure of SDA is identical to that of an L-layered MLNN, and the learned parameters $(\mathbf{W}^l, b^l, c^l)$ can be used to initialize the MLNN. The final layer $\mathbf{h}_i^l$ can then be compared to the desired outputs and the errors can be propagated back to fine-tune the network. SDA learning proceeds layer-wise.

15.2.3.2 Dropout Networks

Unlike SDAs where the corruption is applied stochastically in a layer-wise fashion, the dropout network simply drops a fraction of the network units across *all* layers. Fig. 15.1C shows the dropout learning procedure. During the training stage, in each iteration of the backpropagation, the transformation parameters of a smaller

sub-network are updated. The size of this sub-network is determined by the dropout rate ζ (which, in general, is the same across all layers). The sub-networks across different iterations are chosen randomly, and hence, each transformation parameter is updated approximately $(1 - \zeta)t$ times (where t is the number of gradient iterations). During the test time, the learned parameters are scaled up by $(1 - \zeta)$ corresponding in predicting the final layer representation of the input. One can perform pretraining type initialization in tandem with dropout since the latter does not work layer-wise [33]. In principle, both SDA and dropout are different types of feature corruption based regularization scheme, and the dropout learning procedure has been introduced to address the overfitting issue in MLNNs [33]. However, [41,24] have shown that, in certain cases, pretraining can be 'by-passed' completely when using dropout. This is because the dropout dynamics results in smaller networks, there by reducing the effect of the local minima while also resulting in parameter estimates that are generalizable and robust to input perturbations. With this background we now present the rationale behind our proposed deep learning models for the small sample regime followed by their structure and learning procedure.

15.3 OPTIMAL ENRICHMENT CRITERION

The ideas driving our randomized deep networks are motivated, as discussed earlier, from the design of optimal enrichment criterion. The sample size estimation from (15.1) suggests that, for a given significance α and power $1 - \beta$, the required sample size s (per arm) *increases* as the standard deviation of the outcome σ *increases*, and/or, as the standard change in the outcome δ *decreases*. Recall that the hypothetical RCT presented in Section 15.2.1 asks for the change in outcome to decrease by $(1 - \eta)$ (or the disease should reduce by η). Hence, a smaller η directly indicates that a smaller drug induced change in the treatment is desired and can be detected, there by increasing the robustness of the trial. Clearly, for a fixed number of subjects s, decreasing σ and/or increasing the mean change in the outcome δ, will decrease the detectable drug effect η. Hence, (15.1) implies that one can design an efficient clinical trial by selecting the population and the outcome such that, during the span of the trial, there will be large longitudinal changes in the outcome δ *and* small outcome variance σ^2. Ideally, the trial should *not* include subjects who remain healthy (and/or do not decline) as time progresses because they will reduce the trial's sensitivity for detecting any drug effect. Nevertheless, in general, the trial is always diluted by including those subjects who are unlikely to benefit from the drug, indicating that the alternative hypothesis that the drug induces some effect is moot (for this subgroup). Removing such weak decliners from the trial will result in large δ, but may not necessarily ensure smaller σ^2 for the outcome. To address this, we need to *explicitly* reduce the outcome variance; this is generally not feasible because the outcomes are cognitive and/or blood flow based

measures whose statistical characteristics (range, median, structure, etc.) cannot be altered.

An alternative approach to ensure smaller outcome variance is by designing a *computational* disease marker that predicts the decline in the disease with high confidence. This new marker is explicitly ensured to have smaller variance. If the correlation of this computational marker with the intended trial outcome is strong, then the low variance characteristic of the new marker translates, to a certain extent, to the outcome due to the marker's strong predictability of future decline. Hence, once the new marker is established to have strong correlations to the outcome, one can use it as an "inclusion or enrichment criterion" to filter the trial population – those subjects whose future decline is small according to this enrichment criterion are removed from the trial. An optimal enrichment criterion would then be a computational disease marker with strongest predictability of the disease with smallest possible variance across subjects. Note that the intuition behind this is that reduction of this enricher's variance will indirectly reduce the outcome variance. Using an existing disease marker may not necessarily guarantee this behavior, and hence, one needs to explicitly design such a criterion. In summary, the optimal inclusion criterion should have the following properties:

- strong discrimination power for different stages of the disease – i.e., *no approximation/modeling bias*;
- strong correlation with an existing disease outcomes or other biomarkers – i.e., strong *predictive power* of the disease; and
- small prediction variance – i.e., *small variance* on the intended outcome.

The above requirements can be formulated as a statistical estimation problem. Given the inputs, which may include medical imaging data and/or other relevant types of clinical and/or demographic information, the estimator output is a *new* disease marker satisfying the above requirements. No approximation bias and strong predictive power implies that the estimator needs to be unbiased with respect to the classes/labels. Concurrently with the low prediction variance, the problem of designing the optimal enrichment criterion reduces to constructing a *minimum variance unbiased (MVUB) estimator* of the disease.

15.3.1 ENSEMBLE LEARNING AND RANDOMIZATION

The existence of an MVUB estimator is governed by whether the Cramer–Rao lower bound can be achieved, i.e., any unbiased estimator that achieves this lower bound is referred to as an MVUB [43]. However, finding such an unbiased estimator can be difficult, and especially, in the small-sample regime where $d \gg n$, computing the lower bound itself may is problematic. Instead, in such settings an alternative approach to designing MVUBs is by first generating sets of unbiased estimators that are approximately *uncorrelated* to each other (in the ideal case, independent), and combine them in some reasonable manner to reduce the variance while retaining unbiasedness. This is the classical bootstrap approach to MVUB design in high-dimensional

statistics [44], and the family of models that adapt this approach are broadly referred to as "ensemble learning" methods [45]. The variance reduction behavior follows from ensuring that the estimators/learners have sufficiently small cross-correlation, and so, their linear combination will have smaller variance compared to that of each individual estimator/learner. To see this, let $\mathbf{z}_k$ for $k = 1, \ldots, K$ denote K different random variables (e.g., the outputs from K different estimators/learners), and $\bar{\mathbf{z}}$ denote their mean. Assuming, without any loss of generality, that the variance and cross-covariance of $\mathbf{z}_k$s are σ^2 and ρ, respectively, we have

$$\text{Var}(\mathbf{z}_k) = \sigma^2, \quad \text{Cov}(\mathbf{z}_k, \mathbf{z}_{k'}) = \rho; \quad \text{then} \quad \text{Var}(\bar{\mathbf{z}}) = \frac{\sigma^2}{K} + \frac{(K-1)\rho\sigma^2}{K}. \tag{15.6}$$

Depending on ρ, the variance of $\bar{\mathbf{z}}$ goes from $\frac{\sigma^2}{K}$ to σ^2 (under the assumption that $\rho > 0$, i.e., the estimators are not negatively correlated). In the current setting, since all the K estimators/learners share the same set of input data, they cannot be independent. However, by ensuring that ρ is as small as possible and increasing K, one can make sure that the composed estimator $\bar{\mathbf{z}}$ will have the smallest possible variance. Several approaches may be used to ensure small ρ, most of which are based on *randomly* dividing the input dimensions, data instances and/or other estimation/learning parameters into K subsets and constructing one estimator from each of these subsets [45]. The outputs of these estimators can then be considered to be random (or stochastic, in some sense) approximations of the ideal output with zero-bias and small variance. There is extensive empirical evidence for such strategies where the eventual *ensemble learner* will retain the discriminative power of the individual weak learners while reducing the apparent prediction variance [45]. We will use deep networks to construct an efficient weak learner, followed by presenting a systematic strategy to construct the ensemble – this overall learning procedure will correspond to the randomized deep network model.

15.4 RANDOMIZED DEEP NETWORKS

The ideas from ensemble learning do address the variance reduction requirement for the optimal enrichment criterion. However, the weak learners that go into this ensemble need to be unbiased to begin with. Recall the success of deep learning in learning complex concepts in computer vision, natural language processing and information retrieval [23–25]. It is reasonable to expect that these methods should be translatable to learning problems in medical imaging and neuroimaging with improved performance than the state-of-the-art. This should be plausible, especially because the concepts to be learned in brain imaging might be "less" complex from the perspective of the size of hypotheses spaces to be searched over. Hence, the desired ensemble MVUB could be constructed using deep network weak learners that are trained appropriately to predict the disease. There are a few cavets, however, as described below.

15.4.1 SMALL SAMPLE REGIME AND MULTIPLE MODALITIES

Although the pretraining idea in tandem with the dropout learning address the issue of non-convexity to a certain extent, [46,20,47] have shown extensive evidence that one of the main reasons for the success of deep learning is the availability of large number of unsupervised and/or supervised training instances. The few studies that apply deep learning methods in neuroimaging have reported such observations as well [29,30]. Simply put, the non-convexity, together with the stochasticity that comes from the corruption in DA or the dropout process, demand a very large number of gradient search iterations and data instances to effectively search through the solution space in computing generalizable solutions. As the data dimensionality increases, the dataset size required to ensure that sufficiently many combinations of corrupted/dropped dimensions are passed to the objective also increases proportionally. Now, the fundamental difference between vision type domains, and medical imaging and bioinformatics is the lack of such large datasets. In vision, one has access to extremely large datasets (on the orders of millions of images) – including both large number of unlabeled and supervised instances (e.g., in object recognition, document analysis, and so on). On the other hand, a typical voxel wise imaging study, for instance, will have $n < 500$ subjects while the number of voxels/features (d) will exceed a million – the classical small-sample regime.

Further, in contrast to classical machine learning tasks, the problems in bioinformatics involve data from multiple acquisition types/domains (e.g., brain imaging data including Magnetic Resonance images (MRI), Positron Emission Tomographic (PET) images, several types of cognitive and neuropsychological scores, lists of vascular and blood perfusion data, genetic single nucleotide polymorphisms). Most applications arising in these areas would require efficient statistical models for "fusing" such multi-modal data, especially because they provide distinctive information about the underlying disease. Such multi-modal studies have always been shown to result in higher performance than uni-modal ones [48]. Using classical version of deep architectures, including SDAs or dropout networks, for these multi-modal problems in tandem with $d \gg n$ issues will result in unreliable outputs, with no guarantee of generating either stable or generalizable solutions. This is a direct consequence of under sampling issues in statistical learning (like in VC-dimension or Nyquist sampling analyses), where the number of training instances cannot be below a certain pre-specified number for efficient estimation of the underlying concepts [39], and with presence of multi-modal data the concept is far more complex than from the uni-modal setting.

One of the main contributions of our proposed randomized deep networks is to translate the success of deep learning to multi-modal neuroimaging problems in the small sample regime. We will use neuroimaging terminology like voxels, subjects, etc., to present our models mainly because $d \gg n$ regime is common in brain imaging. The simplest solution to mitigating the $d \gg n$ issue is by pre-selecting the most influential voxels using some statistical test (e.g., t-test based on some known grouping information) and using only these as features within the learning framework downstream. However, this proposal is lossy, in the sense that, non-selected features

are discarded irrespective of how discriminative they are; and so, to ensure least information loss, the processing should be "appropriately" chosen. This not only makes the selection process biased to the task (and not generalizable), but, more importantly, the performance would be entirely driven by the "goodness" of pre-processing. Bourgon et al. [49] discuss this necessity of avoiding bias and influence of data processing on the false positive error rates for the eventual learning task.

On the other hand, one can avoid the feature screening completely by working with slices of the input data (e.g., 2D slices or smaller resolution images from a 3D image). Although this is lossless, working with one slice at-a-time will restrict interactions of voxels and brain regions from anatomically far apart regions. Clearly, allowing for arbitrary far apart interactions in predicting the output label would be reasonable following the hypotheses that, in general, all brain regions have complex biological interactions in generating the final label (e.g., the disease status). An alternative to these extremes is to *categorize* (or *tessellate*) the entire set of voxels into multiple subsets (e.g., spatially contiguous blocks) and learn a network (e.g., SDA or dropout network) on each block separately while allowing for different blocks to interact with each other in some prescribed manner. The network learned in each block will then serve as a weak learner in the above described ensemble (as in boosting [45]). Later, the individual block-wise networks, which interact with each other, can be combined in a meaningful manner yielding a better fit for the dependent variable $\mathbf{y}$.

15.4.2 ROADMAP

The starting point of our proposed models is the rationale that this above described proposal is viable in some sense. Observe that if the number of voxels in each subset/block is comparable to n, and much smaller than the input dimensionality allowed in this smaller subset (denoted by $\tilde{d}$), the learning problem at the level of *this* block is well defined. For instance, d voxels can be divided into B subsets and an SDA or dropout network can be trained on each block separately. On its own, each block's output will not be useful since it corresponds to only a small number of features. But our final output will utilize the outputs from all B blocks, treating each as a stub (i.e., weak learner). This scenario assumes that the tessellation (i.e., the voxel to block mapping) is given or fixed. But in practice, the "correct" tessellation is not known in advance. However, it is possible to marginalize over this as described below. Consider the set of all possible tessellations $\mathcal{C}$ (an exponentially large set). Constructing a learning model, however simple, for each item in $\mathcal{C}$ is unrealistic. Instead, we can use the standard trick of marginalizing over $\mathcal{C}$ by drawing a large number of samples from it to approximate the summand. In other words, by resampling from $\mathcal{C}$ and deriving many possible B (sufficiently large) number of subsets (and learning network weights for them), we can obtain partial invariance to the lack of correct tessellation. Note that one realization of this process corresponds to drawing a sample from $\mathcal{C}$; this process provides *randomization* over clusterings where B different predictions from input instance are combined to generate a single classification/regression output.

The number of blocks B can be fixed ahead of time. The process of assigning voxels to blocks can leverage domain information like neighborhood interactions and/or consistency of correlations across all d dimensions or any other randomized procedure. For instance, if it is known that a local neighborhood has strong interactions, then each such neighborhood can constitute a single block. To allow for arbitrary brain regions to interact with each other possibly with in a block, the block construction should group arbitrary voxels together. Hence, to balance this arbitrary voxel selection and domain information driven block generation, we first 'rank' all the voxels according to some information criterion (e.g., Kullback–Leibler divergence, entropy). We then sample the voxels (for a given block) without replacement according to the cumulative distribution of these ranks. If these blocks are *sufficiently* independent, then any linear combination of their outputs will have smaller variance resulting in an MVUB as desired. However, the input data to all the blocks comes from the same subject, implying that the B weak learners will always be correlated. Nevertheless, we can force them to be approximately uncorrelated by adding another level of randomization over the block generation process. To do this, observe that, beyond the block generation, there are other sets of hyper-parameters corresponding to the learning mechanism of individual blocks like denoising rate, dropout rate, gradient stepsizes, etc. Hence, for a given block we can learn T different number of learning models by randomly generating T different sets of such learning hyper-parameters.

This two-fold randomization will result in an ensemble of $B \times T$ number of weak learners with as small correlation among them as possible. Clearly, increasing the number of weak learners by $T > 1$, decreases the variance of MVUB, while also mitigating the influence of learning parameters. This is the crux of our randomized deep networks. Specific details about the randomization process will be described in the following section as we present the architecture and training mechanisms for these models. Lastly, observe that it is easy to incorporate multi-modal features like MRI or PET within our setup. The simplest way would be to generate blocks from each of the modalities independently, train them separately, and combine their outputs at a later stage. If m denotes the number of modalities, then the blocks from each modality can simply be concatenated resulting in mB number of weak learners. Alternatively, one can construct 'cross-modal' blocks where voxels across multiple modalities are sampled and assigned to a single block. This procedure will have to take into account the model specific tessellating distributions and is far more complex than the intra-modal design mentioned earlier. Since the networks are trained locally on each block, once the blocks are fixed, the voxels within a block do not *directly* interact with those from the other. One can nevertheless construct a feedback procedure that reassigns voxels among blocks based on the goodness of predictions from the previous set of blocks. This feedback is computationally expensive and it may not improve the performance whenever the number and size of the blocks are reasonably large. We refer to the two proposed randomized deep network models as – randomized Denoising Autoencoders (rDA) and randomized Dropout Networks (rDr).

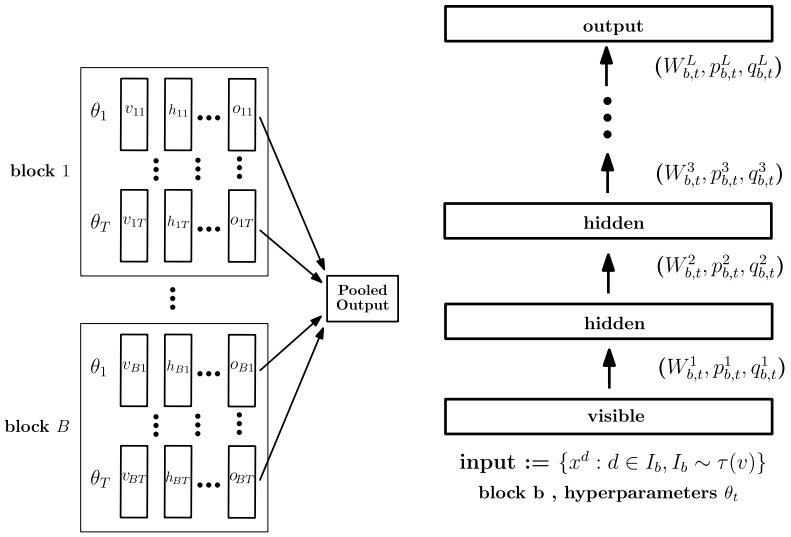

FIGURE 15.2

The architecture of randomized deep network is an ensemble of BT weak learners, where each weak learner is an L-layered MLNN. Each block processes fixed set of voxels s_b of length d_b. Within each block there are T MLNNs.

15.4.3 RDA AND RDR TRAINING

Let $\mathcal{V} = \{v_1, \ldots, v_d\}$ denote the set of voxels. Consider a probability distribution $\tau(v)$ over the voxels $v \in \mathcal{V}$. This is the sampling distribution that governs the block construction, i.e., the assignment of voxels to blocks, and in the simplest case, it is a uniform distribution. For each block $b = 1, \ldots, B$, we sample $d_b \ll d$ (fixed a priori) number of voxels without replacement using the distribution $\tau(v)$. We call this set of voxels, s_b. Each block is presented with T different sets of learning parameters (i.e., denoising rate, gradient learning rate and so on) denoted by $\theta_t \in \Theta$ for $t = 1, \ldots, T$, where Θ is the given hyper-parameter space. This means that each sample from the hyper-parameter space yields one weak learner, i.e., one SDA or Dropout network for one block. Hence, a total of $B \times T$ number of weak learners are constructed. If $\tau(\cdot)$ is uniform, then asymptotically we expect to see one voxel in at least one of the B blocks. As discussed earlier, instead of uniform $\tau(\cdot)$, alternative choices may be used depending on prior information about the importance of including a particular voxel in one/more blocks. Depending on $\tau(\cdot)$ the blocks may be mutually exclusive. The influence of model hyper parameters including B, T, the number of voxels per block d_b, the sampling distribution $\tau(\cdot)$, and the robustness of the model to these choices are described in Section 15.4.3.1.

Fig. 15.2 (left) shows the architecture of randomized deep network with B blocks, each with T weak learners, where each weak learner corresponds to an L-layered

MLNN (as shown in Fig. 15.2 (right)). The outputs from the ensemble of $B \times T$ networks are combined using ridge regression. Algorithm 15.1 summarizes the training procedure. Given training data $(\mathbf{x}_i, \mathbf{y}_i)$ for $i \in \{1, \ldots, n\}$, we first learn the unknown transformations – $(\mathbf{W}^l_{b,t}, p^l_{b,t}, q^l_{b,t}), \forall b \in \{1, \ldots, B\}; t \in \{1, \ldots, T\}; l \in \{1, \ldots, L\}$. Depending on whether the weak learner is an SDA or a dropout network, the learning process for these $B \times T \times L$ transformations will follow the minimization of (15.4) and (15.5), or the discussion from Section 15.2.3.2, respectively. The *Reweigh*$(\cdot)$ operation in Algorithm 15.1 skews the sampling distribution to ensure that the un-sampled voxels (i.e., voxels that are not assigned to any block, yet), will be given priority in the later blocks, while avoiding oversampling of the same set of voxels across multiple blocks.

Concatenating the Lth layer outputs from $B \times T$ learners, we get $\mathbf{H}_i = [[\mathbf{h}^L_{b,t}]]^{B,T}_{1,1}$. The weighted regression pooling then composes these outputs

$$\mathbf{U} \leftarrow (\mathbf{H}^T\mathbf{H} + \lambda\mathbf{I})^{-1}\mathbf{H}^T\mathbf{Y}; \quad \mathbf{H} = [[\mathbf{H}_i]]^n_1; \quad \mathbf{Y} = \left[\left[\mathbf{y}_i\right]\right]^n_1, \tag{15.7}$$

where $\mathbf{U}$ are the regression coefficients, λ is the regularization constant and $\mathbf{Y} = [\mathbf{y}_i]^n_1$. Since the networks are already capable of learning complex concepts (see discussion from Section 15.2.2), the pooling operations that we used was a simple linear combination of the $B \times T$ outputs with ℓ_2-loss providing minimum mean squared error. Clearly, instead of regression or any other fancier combinations, a simple mean of the outputs may also suffice to generate the final output because the networks are already ensured to be as uncorrelated as possible, and so, their mean output is a reasonable estimate of the predicted labels. Observe that the training algorithm and the design of randomized deep networks in general are agnostic to the type of architectures that were used as weak learners. Once the randomized network is trained, its prediction on a new test instance/example is given by (the scaling up of transformations in dropout network case are not shown here to avoid notational clutter, see [33] for more details),

$$\hat{\mathbf{y}} = \mathbf{h}\mathbf{U}; \quad \mathbf{h} = [[\mathbf{z}_{b,t}\mathbf{h}^L_{b,t}]]^{B,T}_{1,1}; \quad \mathbf{h}^l_{b,t} = \sigma(\mathbf{W}^l_{b,t}\mathbf{h}^{l-1}_{b,t} + p^l_{b,t}); \quad \mathbf{h}^0_{b,t} = \mathbf{x}. \tag{15.8}$$

Algorithm 15.1 Randomized deep networks – blocks training

Input: $\theta_t \sim \Theta, \mathcal{V}, B, s_B, L, T, \mathcal{D} \sim \{\mathbf{x}_i, \mathbf{y}_i\}^n_1, \lambda$
Output: $(\mathbf{W}^l_{b,t}, p^l_{b,t}, q^l_{b,t})$
  for $b = 1, \ldots, B$ **do**
    $I_b \sim \tau(\cdot)$
    for $t = 1, \ldots, T$ **do**
      $(\mathbf{W}^l_{b,l}, p^l_{b,l}, q^l_{b,l}) \leftarrow \mathcal{Z}.(\mathcal{D}, L, I_b, \theta_t)$
    end for
    $\tau(\mathcal{V}) \leftarrow$ *Reweigh*$(\tau(\mathcal{V}), I_b)$
  end for

15.4.3.1 Hyperparameters

The hyperparameters of randomized deep network include:

- B, the number of blocks
- T, the number of hyperparameter sets; depth of the network, gradient learning rate, denoising (for rDA) or dropout (for rDr) rates (or other appropriate regularization criteria depending on the weak learning mechanism), number of gradient iterations, etc.
- $\tau(\cdot)$, the sampling distribution over all the voxels for constructing the blocks,
- d_b, the number of voxels within each block (or length of the input layer for weak learners)
- λ, the regularization parameter for ridge regression

Observe that, within each block, the randomization is over the T sets of learning parameters, and hence if the hyperparameter space is sampled uniformly, the model's outputs will be robust to changes in T. The simplest choice for the block-wise sampler $\tau(\cdot)$ assigns uniform probability over all dimensions/voxels as described above. However, we can assign large weights on local neighborhoods which are more sensitive to the disease progression, if desired. We can also setup $\tau(\cdot)$ based on entropy or the result of a hypothesis test. More precisely, entropic measures (like Kullback–Leibler divergence), t-scores or z-scores can be used to estimate the discrimination power of each voxel (this would correspond to performing V number of hypothesis tests; uncorrected). The resulting scores (being positive) can then be normalized and used as the sampling distribution $\tau(\cdot)$. The *Reweigh*$(\cdot)$ step in Algorithm 15.1 takes care of previously unsampled dimensions/voxels. The simplest such re-weighting includes removing the already sampled voxels from $\tau(\cdot)$ (which is the same as sampling voxels without replacement). Although there is no analytic way to setup B and d_b (for $b = 1, \ldots, B$), a reasonably large number of blocks with $d_b = d/B$ would suffice (refer to discussion in Section 15.5 about the choices made in our experiments). The influence of dropout and denoising rates on the performance of deep networks have been well studied empirically [32,33], and [41] analyze the dynamics of the dropout networks as the dropout rate changes. Since such studies already provide ample evidence for setting these rates, we do not explicitly analyze them for our setting.

15.4.4 THE DISEASE MARKERS – RDAM AND RDRM

The randomized deep network from Fig. 15.2 and Algorithm 15.1 will now be adapted to the problem of designing an MVUB of disease spectrum (refer back to our discussion from Section 15.3). Depending on the choice of architectures used, these MVUBs will be referred to as randomized Denoising Autoencoder marker (rDAm) or randomized Dropout network marker (rDrm). The sigmoid nonlinearity ensures that the outputs of individual blocks lie in $[0, 1]$, and so the predictions from the rDA and rNr models on new test examples are bounded between 0 and 1. This is clearly advantageous since a bounded predictor would implicitly guarantee bounded

variance (a desirable property from the perspective of MVUB design; refer to (15.6) and Section 15.3.1) while also covering the entire disease spectrum. Hence, we only need to ensure that each individual block is an unbiased estimate of the disease. We then train the weak learners (the blocks) *only* using healthy controls and completely diseased subjects each labeled $y = 1$ and $y = 0$, respectively. Here the diseased subjects would be those with clinical AD, and all subjects in other stages of the disease (like early or late MCI [9–11]) are not going to be used for training. Since the pooling operation corresponds to a regression (see (15.8)), the test time predictions are the desired disease markers rDAm or rDrm. They will correspond to the *confidence* of rDA or rDr in predicting the subject's decline – closer to 1 if the subject is healthy, or 0 otherwise. Clearly, changing the training setup from regression to classification will modify the interpretation of these predicted markers, but their properties like bounded variance and MVUB would still remain the same.

Recall the discussion in Section 15.2.1 about the sample enrichment procedure where the inclusion criterion decides whether the subject needs to be enrolled in the trial or not. As noted in Section 15.1, from the practical perspective, the inclusion criterion should filter subjects at the trial start point itself. Specifically, the inclusion criterion, either rDAm or rDrm computed at the *baseline* (or trial start point) will then be used to enroll the subjects. Since the markers are bounded between 0 and 1, the enrichment is driven by choosing the "appropriate" *threshold* or cut-off to retain subjects accordingly. This procedure is summarized here:

1. The first check for performing this baseline sample enrichment is to ensure that these markers, computed at the baseline, have strong correlation (or dependence) with other disease markers, some of which would indeed be the intended trial outcomes [50,51]. If the dependencies turn out to be significant, this is evidence of convergent validity, and using baseline rDAm or rDrm as inclusion criteria for enriching the trial population is, at minimum, meaningful.
2. Once this is the case, using the enrichment threshold t $(0 < t < 1)$, the enriched cohort would include only those subjects whose baseline rDAm or rDrm is smaller than t (closer to being diseased). Alternatively, by avoiding to choose the optimal cut-off t, one can instead include a fixed fraction (e.g., 1/4th or 1/3rd) of the whole population whose baseline rDAm or rDrm is closest to 0 (fixing a fraction automatically fixes t).

Clearly, the choice of t is vital here, and one way to choose it is by comparing the mean longitudinal change of some disease markers (MMSE, CDR, and so on) for the enriched cohort as t goes from 0 to 1. The optimal t would correspond to a discontinuity or a "bump" in the change trends as t increases. We discuss more on these issues as we present evaluate rDA and rDr.

15.5 EXPERIMENTS

15.5.1 PARTICIPANT DATA AND PREPROCESSING

Imaging data including [F-18]Florbetapir amyloid PET (AV45) singular uptake value ratios (SUVR), FDG PET SUVRs and gray matter tissue probability maps derived from T1-weighted MRI data, and several neuropsychological measures and CSF values from 516 individuals enrolled in Alzheimer's Disease Neuroimaging Initiative-II (ADNI2)[1] were used in our evaluations. Of these 516 persons (age 72.46 ± 6.8, female 38%), 101 were classified as AD (age 75.5 ± 5.1), 148 as healthy controls (age 70.75 ± 7), and 131 and 136 as early and late MCI (age 74.3 ± 7.1 and 75.9 ± 7.7), respectively, at baseline. There was a significant age difference across the four groups with $F > 10$ and $p < 0.001$. Among the MCIs, 174 had positive family history (FH) for dementia, and 141 had at least one Apolipoprotein E (APOE) e4 allele. CSF measures were only available at baseline, and three time point data (baseline, 12 months, and 24 months) was used for the rest. The imaging protocols follow the standards put forth by ADNI. MRI images are MP-RAGE/IR-SPGR from a $3T$ scanner. PET images are 3D scans consisting of four 5-min frames[2] from 50 to 70 min post-injection for [F-18]Florbetapir PET, and six 5-min frames from 30 to 60 min post injection for FDG PET. Modulated gray matter tissue probability maps were segmented from the T1-weighted MRI images (other tissue maps are not used in our experiments) using voxel-based morphometry [52]. The segmented map was then normalized to Montreal Neurological Institute (MNI) space, smoothed using 8 mm Gaussian kernel, and the resulting map was thresholded at 0.25 to compute the final gray matter image. All PET images were first co-registered to the corresponding T1 images, and then normalized to the MNI space. Manually constructed masks of pons, vermis, and cerebellum were then used to scale these PET maps by the average intensities in pons and vermis (FDG PET SUVR) and cerebellum (Florbetapir PET SUVR). All preprocessing was done in SPM8 [53].

15.5.2 EVALUATIONS SETUP

We train the randomized deep networks using only baseline imaging data from all the three modalities, MRI, FDG PET, and AV45 PET with diseased (AD, labeled 0) and healthy (CN, cognitively normal, labeled 1) subjects. When testing on MCI subjects, these trained models output a multi-modal rDAm and rDrm, which represent the confidence of rDA and rDr that a given MCI subject is (or is not) likely to decline. We only use baseline imaging data for training, there by making the models deployable in practice, while the predictions can be performed on MCIs at baseline and/or future time-points. $B = 5000$, $d_b = d/B$ (i.e., each voxel appears in only one block), and $\lambda = 1$ for both rDA and rDr. $\tau(\cdot)$ is based on differentiating ADs and CNs using KL divergence (refer to Section 15.4.3.1). Multiple combinations of B, d_b and $\tau(\cdot)$ (including uniform and t-score based) were also evaluated, however, none of them gave any significant improvements over the above settings. The blocks construction for multi-modal rDA and rDr did not use cross-modal sampling to ensure manage-

able computational burden (see Section 15.4.2). Within this setup, our evaluations are three-fold. Since the test data are MCIs, we first evaluate if baseline rDAm and rDrm differentiate early MCI from late MCI, and parental family history as a contributing risk factor. These evaluations include the baseline markers derived from seven different combinations corresponding to the three imaging modalities available.

After checking the construct that multi-modal markers are superior to unimodal markers, we evaluate the premise whether the multi-modal rDA and rDr markers are good disease progression markers. We demonstrate this by computing the dependence of these multi-modal baseline rDAm and rDrm with well-known outcome measures including, Mini Mental State Examination (MMSE), Alzheimer's Disease Assessment Scale (ADAS Cognition 13), Montreal Cognitive Assessment (MOCA), Rey Auditory Verbal Learning Test (RAVLT), neuropsychological summary score for Memory (PsyMEM), summary score for Executive Function (PsyEF), hippocampal volume from gray matter images, Clinical Dementia Rating sum of boxes (CDR-SB), a binary marker for conversion from MCI to AD (DxConv), CSF levels including Tau τ, Phospho-Tau $p\tau$, Amyloid Beta $A\beta 42$, ratios of τ and $p\tau$ with $A\beta 42$, APOE allele 4 and maternal/paternal family history (FH). Please see [54,10,51] and other appropriate references[3] therein for complete details about these disease and at-risk markers. For continuous markers, we used the Spearman Rank Order Correlation coefficient to assess the dependencies and accepted those statistics as significant whenever the corresponding $p < 0.05$. For binary markers the t-test was used with the same significance value. Observe that we are interested in evaluating the predictive power of baseline rDAm and rDrm. Specifically, we report the correlations of baseline rDAm with these markers at 12 months and 24 months, and also the longitudinal change with this one year, thereby providing evidence that whenever the baseline markers are closer to 0, the subject's longitudinal changes are, in fact, steeper/stronger.

Once the correlation construct is appropriately validated, we evaluate the use of baseline rDAm and rDrm for sample enrichment. We compute the sample sizes (using (15.1) based on the discussion in Section 15.2.1) required when using the above cognitive, neuropsychological, diagnostic and other imaging-based outcome measures with (and without) rDAm or rDrm based enrichment. The sample size trends are computed across different enrichment thresholds (see Section 15.4.4). For better interpretation of the estimates from the perspective of a practitioner or clinician, we estimate the effect size as a function of the marker enrichment cut-off for a given (fixed) sample size. We also compute the performance improvement from using our markers relative to the alternative imaging-derived enrichers including ROI summaries from FDG and florbetapir images[4] with particular attention to the current state-of-the-art imaging based computational summary measure that we refer to as a Multi-Kernel Learning marker [4]. MKLm is based on a Multi-Kernel support vector machine (MKL) [4] that tries to harmonize contributions from multiple imaging modalities for deriving a maximum margin classifier in the concatenated Hilbert spaces. Unlike traditional support vector machines, MKLm uses a linear combination of kernels and solves for both the weights on the kernels as well as the normal to the

Table 15.1 rDA and rDr vs. MKLm. A, amyloid; F, FDG; and T, T1GM

Model	Amyloid	FDG	T1GM	A+F	A+T	F+T	A+F+T
(A) Early versus late MCI							
MKL	$20.5^{\dagger}$	$16.8^{\dagger}$	$16.5^{\dagger}$	$16.4^{\dagger}$	$20.4^{\dagger}$	23.6*	27.9*
rDA	22.1*	$9.7^{\dagger}$	$20.0^{\dagger}$	$19.5^{\dagger}$	24.1*	21.2*	27.6*
rDr	$20.1^{\dagger}$	$11.5^{\dagger}$	$20.3^{\dagger}$	$17.5^{\dagger}$	23.0*	21.2*	26.9*
(B) Family history: positive versus negative							
MKL	4.3**	$7.5^{\dagger}$	5.3**	$7.3^{\ddagger}$	$6.8^{\ddagger}$	6.6**	$8.3^{\ddagger}$
rDA	4.7**	$11.8^{\dagger}$	$11.2^{\dagger}$	$6.8^{\dagger}$	$12.4^{\dagger}$	$13.2^{\dagger}$	$13.3^{\dagger}$
rDA	4.6**	$9.9^{\dagger}$	$12.0^{\dagger}$	$6.9^{\dagger}$	$11.7^{\dagger}$	$13.2^{\dagger}$	$12.0^{\dagger}$

hyper-plane concurrently [4]. Similar to the proposed models, MKL is trained using AD and CN subjects, and the corresponding predictions on MCIs are referred to as MKL markers (MKLm). Note that whenever the labels correspond to continuous predictors instead of class indices (like AD vs. CN), we use ϵ-support vector machine version of MKL.

15.5.3 RESULTS

Table 15.1 shows the discrimination power of rDA and rDr for classifying early and late MCI (Table 15.1A) and family history (Table 15.1B). Both rDA and rDr perform better than the baseline MKL. Tables 15.2 and 15.3 correspond to the predictive power of baseline rDAm and rDrm, respectively. They show the Spearman correlations and t-statistics of the baseline multi-modal markers with cross-sectional scores (baseline, 12 and 24 months) and longitudinal change (12 and 24 months) in other disease markers. Negative correlations indicate that the corresponding measures (ADAS, τ, $p\tau$, $\tau/A\beta 42$, $p\tau/A\beta 42$) increase with disease progression. Across both rDAm and rDrm, large correlations with $r > 0.45$ and/or $p < 10^{-4}$ (denoted by the superscript *), were observed with baseline summary measures (see the second column in Tables 15.2 and 15.3), specifically with ADAS, neuropsychological (memory and executive function) composite scores, hippocampal volume, and CSF levels involving $A\beta 42$. For both rDA and rDr, FH had a smaller influence on baseline rDAm compared to APOE. All the cross-sectional correlations (columns 2–4, Tables 15.2 and 15.3) were significant ($p < 10^{-4}$). The correlations of baseline markers with longitudinal change (last two columns) were significant with $r > 0.21$ and $p < 0.001$ for all the measures (except few cases involving PsyEF and MOCA). rDAm's correlations are stronger than that of rDrm's across all the constructs. Tables 15.1–15.3 provide strong evidence for both rDAm's and rDrm's disease predictive power.

Table 15.4, Figs. 15.3 and 15.4 show the relevance of rDA and rDr for enrichment. Table 15.4 shows the coefficient of variation (CV, the ratio of standard deviation to mean) of rDAm and rDrm for three different populations of interest – all MCIs, late MCIs and MCIs with positive FH. The proposed markers' CV is smaller than MKLm for all three populations and combinations of modalities – making it a better candi-

Table 15.2 Correlations of multi-modal baseline rDAm

Biomarker	Baseline	Cross-sectional		Longitudinal change	
MMSE	0.39*	0.49*	0.45*	0.21†	0.33‡
ADAS	−0.56*	−0.58*	−0.53*	0.21†	−0.53‡
MOCA	0.48*	0.51*	0.59*	0.06	0.59*
RAVLT	0.49*	0.52*	0.57*	0.13**	0.57†
PsyMEM	0.56*	0.57*	0.59*	0.28*	0.42†
PsyEF	0.52*	0.57*	0.46*	0.15**	0.26**
HippoVol	0.72*	0.74*	0.79*	0.33*	0.47*
CDR-SB	−0.33*	−0.49*	−0.55*	−0.36*	−0.53*
DxConv	—	21*	31*	21*	31*
τ	−0.39*	—	—	—	—
$p\tau$	−0.40*	—	—	—	—
$A\beta 42$	0.55*	—	—	—	—
$\tau/A\beta 42$	−0.52*	—	—	—	—
$p\tau/A\beta 42$	−0.52*	—	—	—	—
APOE	3.47†	—	—	—	—
FH	2.16**	—	—	—	—

Table 15.3 Correlations of multi-modal baseline rDrm

Biomarker	Baseline	Cross-sectional		Longitudinal change	
MMSE	0.37*	0.41*	0.31*	0.21*	0.25‡
ADAS	−0.44*	−0.42*	−0.34*	−0.22*	−0.18**
MOCA	0.41*	0.35*	0.33*	0.09	0.29‡
RAVLT	0.43*	0.34*	0.27‡	0.11	0.20‡
PsyMEM	0.48*	0.39*	0.34‡	0.22*	0.29‡
PsyEF	0.45*	0.38*	0.22	0.13**	0.10
HippoVol	0.62*	0.50*	0.25†	0.22†	0.15**
CDR-SB	−0.33*	−0.35*	−0.35*	−0.30*	−0.31*
DxConv	—	17*	17*	11†	10†
τ	−0.34*	—	—	—	—
$p\tau$	−0.35*	—	—	—	—
$A\beta 42$	0.50*	—	—	—	—
$\tau/A\beta 42$	−0.46*	—	—	—	—
$p\tau/A\beta 42$	−0.46*	—	—	—	—
APOE	6.9‡	—	—	—	—
FH	5.01**	—	—	—	—

date to be used as a prediction measure as well as an enricher (refer to (15.1), where the right hand side includes terms depending on CV^2). CV of rDAm is smaller than rDrm. Each plot in Figs. 15.3 and 15.4 corresponds to the mean longitudinal change of some disease marker, *after* the MCI population (the test set) is enriched by removing the weak decliners (subjects with baseline rDAm or rDrm above a certain

Table 15.4 CV of rDAm and rDrm vs. MKLm

Modality	Marker	MCIs	LMCIs	FHMCIs
Amyloid	MKLm	0.56	0.70	0.42
	rDAm	0.49	0.57	0.41
	rDrm	0.55	0.60	0.46
FDG	MKLm	0.49	0.53	0.39
	rDAm	0.33	0.36	0.26
	rDrm	0.45	0.44	0.30
T1MRI	MKLm	0.55	0.60	0.48
	rDAm	0.36	0.42	0.26
	rDrm	0.41	0.42	0.26
A+F	MKLm	0.52	0.63	0.39
	rDAm	0.42	0.49	0.33
	rDrm	0.50	0.57	0.39
A+T	MKLm	0.56	0.67	0.42
	rDAm	0.41	0.49	0.29
	rDrm	0.50	0.51	0.29
F+T	MKLm	0.51	0.58	0.41
	rDAm	0.34	0.38	0.25
	rDrm	0.41	0.44	0.31
A+F+T	MKLm	0.54	0.65	0.39
	rDAm	0.41	0.50	0.28
	rDrm	0.44	0.57	0.30

cut-off t, shown on the x-axis). For both rDA and rDr, the plots show that MMSE, MOCA, hippocampal volume, CDR-SB, and DxConv have large changes when weak decliners are progressively removed – strong evidence for rDAm's and rDrm's predictive power. Specifically, for some measures the changes are much steeper for 24 months than 12 months (black and blue lines in each plot). PsyEF resulted in irregular changes at different time points for both rDA and rDr.

Tables 15.5–15.8 show sample sizes when multi-modal baseline rDAm and rDrm are used as enrichers, at 80% statistical power and 0.05 significance level, with a hypothesized treatment effect of $\eta = 0.25$ (i.e., 25% decrease in the disease). Tables 15.5 and 15.6 show the sample sizes at four different rDAm and rDrm enrichment cut-offs (third to last columns), respectively. Recall that higher values of the markers at baseline imply closer to being healthy. Hence, enrichment entails to filtering out all subjects whose baseline rDAm or rDrm are above some chosen cut-off. Results show that, compared to the no-enrichment regime (second column), the sample estimates from rDAm or rDrm enrichment are significantly smaller, with more than 5 times reduction when using bottom 20% and 25% percentiles (third and fourth columns).

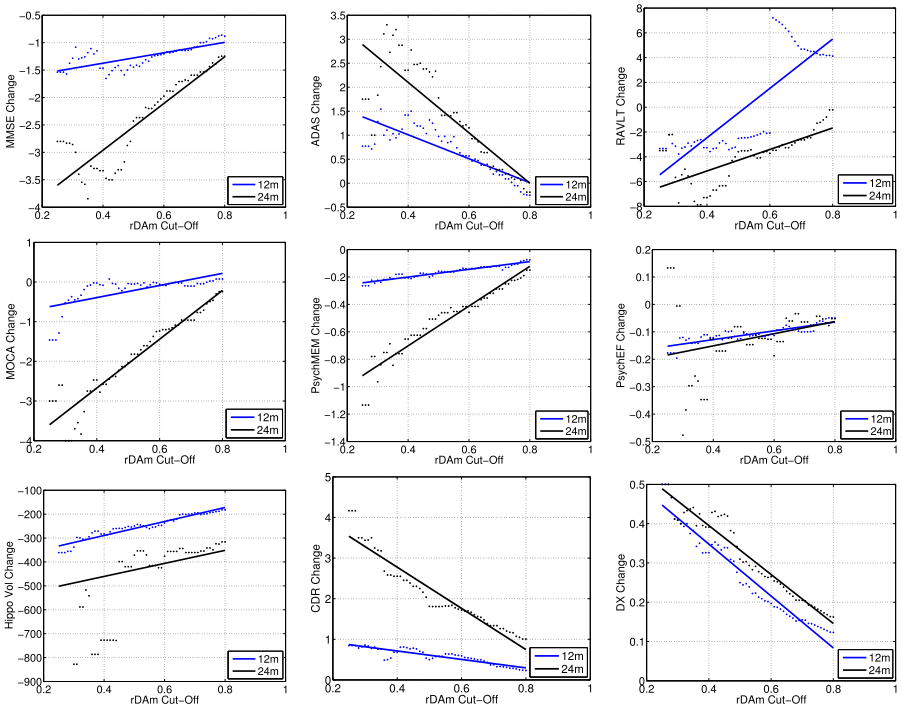

FIGURE 15.3

Longitudinal change of measures vs. multi-modal baseline rDAm.

Table 15.5 Sample sizes with multi-modal baseline rDAm enrichment

Outcome measure	No enrichment	Bottom 20% rDAm ≤ 0.41	Bottom 25% rDAm ≤ 0.46	Bottom 33% rDAm ≤ 0.52	Bottom 50% rDAm ≤ 0.65
MMSE	1367	200	239	371	566
ADAS	>2000	775	945	>2000	>2000
MOCA	>2000	449	674	960	1919
RAVLT	>2000	591	1211	>2000	>2000
PsyMEM	>2000	420	690	786	1164
PsyEF	>2000	>2000	>2000	>2000	>2000
HippoVol	>2000	543	1504	1560	1675
CDR-SB	1586	281	317	430	433
DxConv	895	230	267	352	448

The sample sizes from rDAm enrichment are smaller than those from rDrm, following the previous observation that the CV's of rDrm are larger (refer to Table 15.4). In particular, with rDAm; MMSE, CDR-SB, and DxConv give consistently smaller

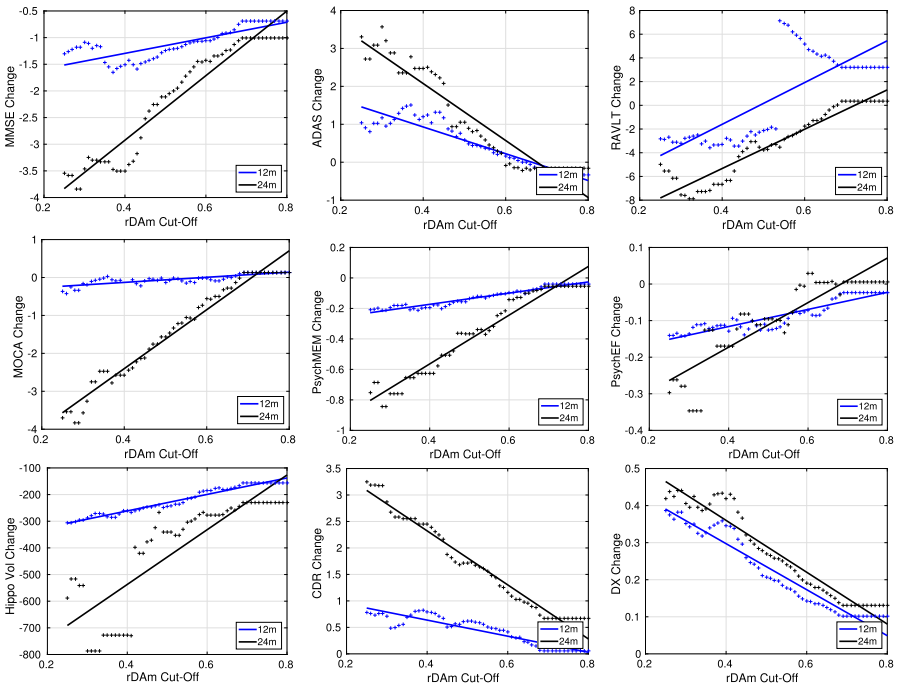

FIGURE 15.4

Longitudinal Change of measures vs. multi-modal baseline rDrm.

Table 15.6 Sample sizes with multi-modal baseline rDrm enrichment

Outcome measure	No enrichment	Bottom 20% rDAm ≤ 0.39	Bottom 25% rDAm ≤ 0.44	Bottom 33% rDAm ≤ 0.58	Bottom 50% rDAm ≤ 0.70
MMSE	1367	252	341	394	560
ADAS	>2000	930	1770	>2000	>2000
MOCA	>2000	655	795	1106	1866
RAVLT	>2000	1556	>2000	>2000	>2000
PsyMEM	>2000	442	700	799	1215
PsyEF	>2000	>2000	>2000	>2000	>2000
HippoVol	>2000	621	1100	1846	>2000
CDR-SB	1586	307	524	608	618
DxConv	895	232	287	307	429

estimates (200 to 600) across all columns (the four different percentiles). ADAS and PsychEF still required very large sizes (774 and >2000, respectively) even at 20% enrichment percentile. Similar trends are observed for rDrm.

Table 15.7 Sample sizes with rDAm + FH and/or APOE enrichment

Outcome measure	No enrichment	FH only	APOE only	rDAm only	rDAm + FH	rDAm + APOE	rDAm + both
MMSE	1367	1668	1015	200	182	240	186
ADAS	>2000	>2000	>2000	775	574	328	271
MOCA	>2000	>2000	>2000	449	516	326	334
RAVLT	>2000	>2000	>2000	591	394	484	332
PsyMEM	>2000	>2000	>2000	420	481	310	333
PsyEF	>2000	>2000	>2000	>2000	>2000	1337	721
HippoVol	>2000	>2000	>2000	428	391	274	246
CDR-SB	1586	1787	763	281	255	217	225
DxConv	895	932	509	230	244	170	192

Table 15.8 Sample sizes with rDrm + FH and/or APOE enrichment

Outcome measure	No enrichment	FH only	APOE only	rDrm only	rDrm + FH	rDrm + APOE	rDrm + both
MMSE	1367	1668	1015	252	292	301	306
ADAS	>2000	>2000	>2000	930	1001	1151	642
MOCA	>2000	>2000	>2000	655	669	636	669
RAVLT	>2000	>2000	>2000	1556	1102	>2000	1544
PsyMEM	>2000	>2000	>2000	442	496	512	385
PsyEF	>2000	>2000	>2000	>2000	>2000	>2000	941
HippoVol	>2000	>2000	>2000	621	799	698	698
CDR-SB	1586	1787	763	307	316	351	392
DxConv	895	932	509	232	259	219	239

Tables 15.5 and 15.6 further enrich the population after using rDrm and rDrm with extra covariate information like FH and/or APOE status. Specifically, here we explicitly filter out those MCI subjects who are *not* FH and/or APOE positive before performing baseline rDAm or rDrm enrichment. Clearly, the sample sizes further decrease because of this extra filtration as shown in last three columns of Tables 15.5 and 15.6. APOE as a covariate resulted in smallest possible estimates in general (<350 per arm with MMSE, CDR-SB, and DxConv outcomes) across all the outcomes except PsyEF (last two columns in Table 3). Interestingly, they are smaller than the sample sizes from using both APOE and FH as covariates (last column). DxConv as an outcome with rDAm or rDrm + APOE enrichment yields a sample size of 170 and 219, respectively. Figs. 15.5 and 15.6 show the detectable effect sizes as rDAm and rDrm enrichment cut-offs are varied for a fixed sample size of 500 per arm. The detectable effect size ($1 - \eta$) decreases as more weak decliners are filtered out. This can be seen by the increase of η (y-axis) as the rDAm or rDrm cut-offs on x-axis decrease, specifically for MMSE, CDR-SB, and DxConv outcomes. These plots are very useful from a clinician's or practitioner's perspective, as will be dis-

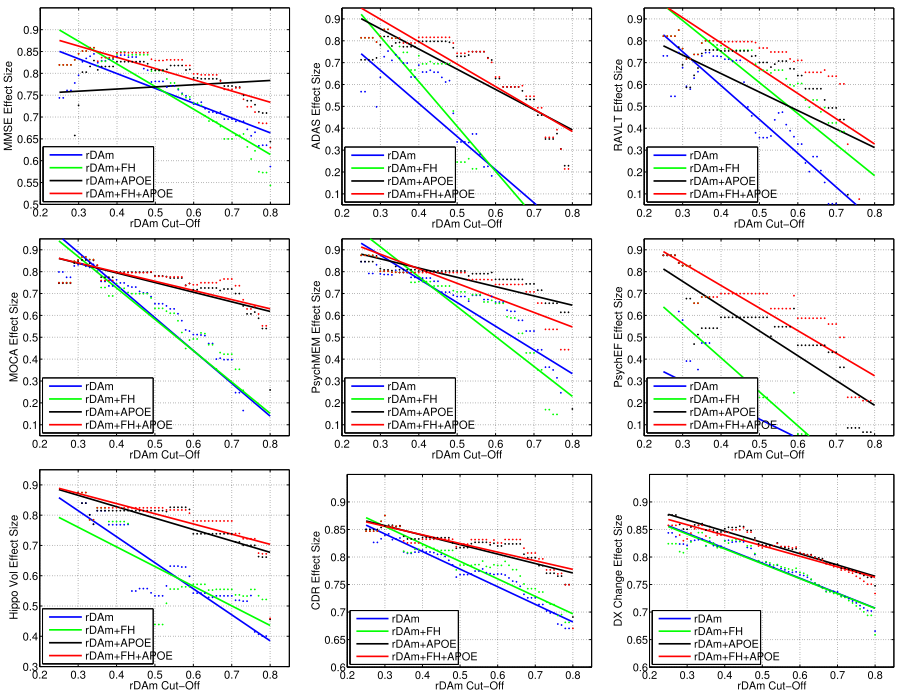

FIGURE 15.5

Effect sizes vs. multi-modal baseline rDAm cut-offs.

cussed later in Section 15.6. Finally, Table 15.9 compares our proposed enrichers with other imaging-derived inclusion criteria (the cut-off for all the enrichers corresponds to including the strongest 20% decliners in their respective scales). Both rDAm and rDrm consistently outperformed other alternatives (and between them rDAm was better), with up to 2 times smaller estimates than MKLm, and much larger reductions compared to uni-modal summaries (hippocampal volume, FDG and AV45 ROIs).

15.6 DISCUSSION

The ability to design clinical trials with smaller sample sizes but sufficient statistical power will enable the implementation of affordable, tractable, and hopefully, conclusive trials. Efficiency is seriously compromised in trials where there is poor biomarker specificity of disease progression and when the outcomes contain relatively high amounts of error variance. Determining whether promising treatments are effective in the MCI phase of AD requires accurate identification and inclusion of only those MCI participants most likely to convert to AD and selection of outcomes that are both disease related and possess optimal measurement properties. We have

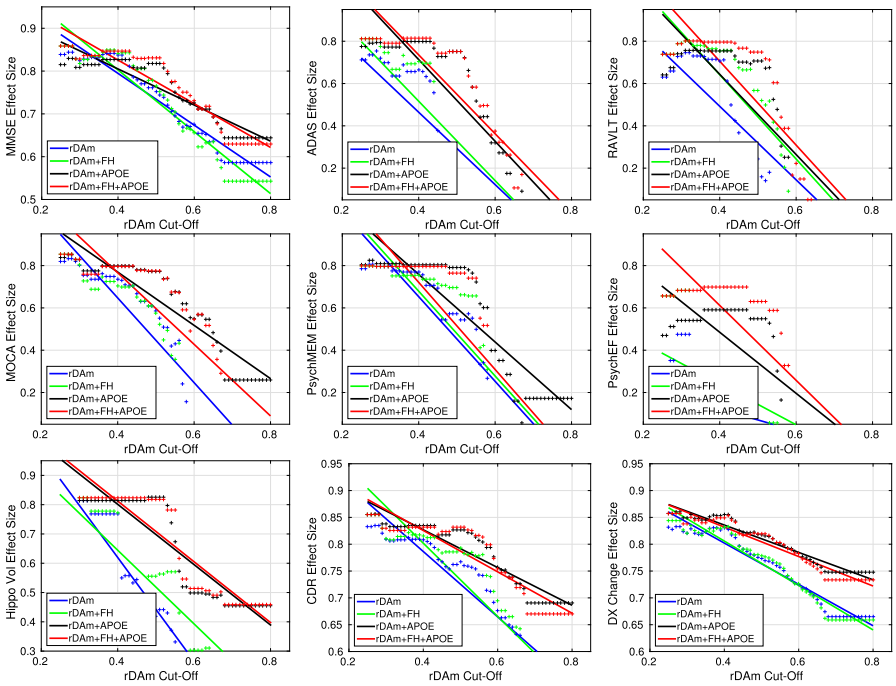

FIGURE 15.6

Effect sizes vs. multi-modal baseline rDrm cut-offs.

Table 15.9 Multi-modal baseline rDAm and rDrm vs. other enrichers

Sample enricher	Outcome measure							
	MMSE	ADAS	MOCA	RAVLT	PsyMEM	HippoVol	CDR-SB	DxConv
HippoVol	540	>2000	1005	1606	1009	>2000	389	420
FDG	384	1954	579	>2000	832	752	415	371
AV45	224	>2000	875	>2000	826	698	382	443
FDH	296	>2000	705	>2000	826	722	397	402
MKLm	228	874	827	896	487	877	295	284
rDAm	200	775	449	591	420	543	281	230
rDrm	252	930	655	1556	442	621	524	287

shown that the sample size required to detect a treatment effect can be substantially reduced using the proposed inclusion strategy. The central message of our empirical evaluations is that the multi-modal markers based on our proposed randomized deep network leaning models have good predictive power in identifying future disease progression, as shown in Tables 15.1–15.3 and Figs. 15.3–15.4. Together with the rDA's or rDr's capacity to reduce prediction variance (Table 15.4), we see smaller sample estimates compared to existing imaging-derived enrichers, as shown in Table 15.9, across many trial outcomes.

Tables 15.1–15.3 support the general consensus that imaging data captures disease progression [4,51]. This can be seen from the very strong correlations of baseline rDAm and rDrm with cross-sectional and longitudinal changes in several cognitive scores (last four columns). It should be noted that high correlations with hippocampal volume across all time-points, which is a structural summary from MRI image, are expected because T1 MRI images at baseline is used in the construction of the markers. Although hippocampus voxels are used in rDA and rDr, its inclusion as an outcome in our experiments is primarily for completeness and continuity with existing AD imaging studies, where it has been extensively studied [6,18,50,51]. Interestingly, FH had a lower dependence on the baseline markers which might be because its influence is superseded by actual neurodegeneration once a subject reaches MCI stage (i.e., FH may play a much stronger role in the asymptomatic phase). Note that we did not correct for age (and other covariates like brain volume) because the markers reported in Tables 15.1–15.3 are used directly with no covariate correction in our later evaluations on sample enrichment (Tables 15.5–15.8). This is based on the assumption that an actual RCT would not need to correct for the individual's age to evaluate eligibility and the baseline markers are agnostic to all such variables.

Observe that most classification based measures which are used as computational disease markers are generally unbounded [5]. These include the prediction score from an SVM based classification model on a test subject [4], or summary measures like S-score, t-score, F-score, etc. [6]. Unlike these measures, the proposed markers are bounded between 0 and 1, using which we can visualize its predictive power without any post-hoc normalization (as shown in Figs. 15.3–15.4). Except for PsyEF, all other measures used as outcomes had steeper changes (in Fig. 15.3) over time as baseline rDAm decreased, and in none of the cases was there a clear elbow separating weak and strong decliners. Similar trend is also observed for baseline rDrm. This shows that the disease progression is gradual from healthy to AD, and any classifications (like early and late MCI) are mostly artificial – an observations made earlier in alternate studies on decline [50,54,12]. Nevertheless, it is of clinical interest to analyze different groups of decliners like late and early MCI, and although the baseline MKLm (the current state-of-the-art in AD classification [4]) picks up these group differences as well, the proposed models have higher delineation power (Table 15.1). In particular, the p-values for rDAm and rDrm for FH+ vs. FH− case are an order of magnitude smaller than MKLm. These show that, in terms of classification accuracy, rDA and rDr are at least as good as a current state-of-the-art machine learning derived measures.

It is interesting to see rDAm's and rDrm's high predictive power for DxConv (Tables 15.2–15.3 and Figs. 15.3–15.4), implying that subjects with smaller baseline rDAm (closer to 0) have very high likelihood of converting from MCI to AD, providing additional evidence that both baseline (and multi-modal) rDAm and rDrm are good predictive disease markers; more so, the predictions from randomized deep networks are good disease markers. Beyond the predictive power, the CVs for proposed models are much smaller than MKLm (Table 15.4) – the central argument that motivated the design of randomized deep networks for sample enrichment (refer to

Section 15.3). Using CV as a surrogate for prediction variance gives interesting inferences about the stages of the disease, for instance, the CVs for MCIs with FH+ are smaller than that of late MCIs (from Table 15.4). This suggests that a significant number of late MCIs currently have only a mild dementia in terms of both rDAm and rDrm. Most of the prediction power and sample size experiments focused on the multi-modal rDAm and rDrm (using all three modalities – amyloid and FDG PET and T1 MRI) since several existing studies including [4,6,2,55] and many others, and the performance results in Table 15.1, have shown the non-trivial benefit of multi-modal disease markers.

Since the proposed markers are lower bounded to 0 and no elbows are seen in Figs. 15.3–15.4, there is no phase change, and we can always select a fixed fraction of subjects that are closest to 0 on the baseline rDAm or rDrm scales, and claim that they are the strong decliners that should be included in a trial. The exact value of such fraction would depend on the logistics and size of the intended trial. This is the reason for the bottom-fraction based enrichment using multi-modal baseline markers as shown in Tables 15.5–15.9. Further, the high predictive power of baseline rDAm and rDrm solves an important bottleneck with existing approaches to designing inclusion criteria that use longitudinal data (e.g., tensor-based morphometry) [14,5]. Deploying such methods in practice implies that the trial screening time should be at least a year or longer, which is not practical (both in terms of cost involved and other logistics). Although longitudinal signals are much stronger than cross-sectional ones, the results in Tables 15.1–15.3 and Figs. 15.3–15.4 show that the randomized deep network based markers at trial start-point can still be used with no loss of information, saving trial resources and reducing the cost of trial setup.

A broad observation across Tables 15.5–15.9 is that baseline rDAm is better than rDrm with smaller sample sizes overall, although the trends are the same for both. Although both SDA and dropout network work with feature denoising, rDA seems to be clearly better at disease prediction as well (Tables 15.1–15.3). This higher sample estimates for rDr might be driven by its higher prediction variance (Table 15.4). Few reasons for this broad trend are discussed here. First observe that rDr lacks the unsupervised pretraining step unlike rDA (refer to Sections 15.2.3.2 and 15.4.3). Secondly, higher dropout rates might be "unsuitable" for brain images unlike vision or other machine learning datasets. To see this, observe that although complex interactions across voxels/dimensions are common in brain images, they are, nevertheless, registered to a common coordinate space (i.e., unlike object recognition or categorization data, a specific set of voxels in rDA and rDr always correspond to a specific region). In this registered space, the signals on brain images (voxel intensities) are, in general, very weak, and subtle changes in them will correspond to a disease signature [56,4,57,54]. For example, the disease signature in hippocampal region corresponds to loss (or dampening) of voxel intensities in a certain manner. Large dropout rates corrupts the images drastically resulting in loss of signal, and in the worst case, the dropped network from healthy subject might *look very similar to* being a diseased one. Unlike rDr, the post-hoc fine tuning in rDA (which does not involve corruption) compensates for these issues, thereby resulting in better performance. Hence,

for the rest of the discussion, we focus mainly on the sample sizes estimated from multi-modal baseline rDAm enrichment instead of rDrm.

MMSE, CDR-SB, and DxConv sample estimates (in Tables 15.5–15.8) outperform all other alternate outcomes considered here, even in the no-enrichment regime. This may be counter intuitive because of the simplicity of MMSE compared to other composite scores like PsyMEM and PsyEF (neuro-psych memory and executive function composites). It is possible that the composite nature of these measures increases the outcome variance, and thereby increases the sample estimates when used as trial outcomes. Since our population is entirely MCIs, it may be expected that the distribution of baseline rDAms is fairly uniform between 0 and 1, but is not the case as shown from rDAm enrichment cut-offs at each of the four percentiles considered (the top row of last four columns in Table 15.5). More precisely, the bottom 50% corresponds to a cut-off of 0.65, and 33% corresponds to 0.52, which indicates that more than two-thirds of MCIs in the ADNI2 cohort are healthier (i.e., weak decliners), and also that enrichment is important. This idea has also been identified by others using cognitive characteristics [58]. Ideally, we expect to observe a particular baseline rDAm cut-off (an elbow) at which there might be the highest decrease in estimates for all outcomes in Tables 15.5 and 15.7. The elbow should be a natural threshold point that separates strong and weak decliners on baseline rDAm scale. However, the trends in the last four columns do not seem to suggest such a threshold, which is not surprising following Fig. 15.3 and the corresponding discussion above. Specifically, ADAS and RAVLT seem to have an elbow between 25% and 33%, while for MMSE, CDR-SB, and DxConv, the elbow is beyond 50%.

Covariate information, or rather, a preliminary selection based on a factor like FH, is almost always helpful in estimating group effects (Tables 15.7–15.8). It has been observed that subjects with positive FH (either maternal or paternal) and/or APOE e4 positive may have stronger characteristics of dementia [59]. This implies that, instead of starting off with all MCIs, it is reasonable to include only those MCIs with positive FH and/or positive APOE e4, and then perform the baseline rDAm or rDrm enrichment on this smaller cohort. APOE had a higher dependence on both rDAm and rDrm compared to FH (from Tables 15.2–15.3), which resulted in a smaller sample sizes when using APOE or APOE + FH in tandem with rDAm enrichment (last two columns), than using rDAm + FH (sixth column) for all the cases except MMSE (row 1 in Table 15.7). Note that Table 15.7 corresponds to bottom 20% baseline rDAm enrichment of which about half were FH and/or APOE positive. The strong performance of DxConv with small sample sizes may be because it summarizes the conversion of MCI to AD using longitudinal information, where as rDAm tries to predict this conversion using baseline information alone. We note that, since we have 267 MCIs to begin with, even with rDAm enrichment alone, a bottom 20% enrichment (third column, Tables 15.5 and 15.6) corresponds to a population size of 52, implying that the estimates might be noisy.

Overall, Tables 15.5 and 15.7 support the efficacy of rDAm enrichment (similar trends are seen for rDrm enrichment from Tables 15.6 and 15.8); however, an interesting way to evaluate the strength of rDAm is by fixing the number of trial-enrolled

subjects and computing the detectable treatment size (η). If in fact, baseline rDAm successfully selects strong decliners, then the trial should be able to detect smaller expected decrease in disease (i.e., smaller $1-\eta$ or larger η, refer to (15.1)). Fig. 15.5 shows exactly this behavior for rDAm (and Fig. 15.6 for rDrm), where η (y-axis) increases drastically as rDAm cut-offs (x-axis) are decreased (especially for MMSE, CDR-SB and DxConv). From the perspective of a practitioner, such plots are very useful. Specifically, they give tools for evaluating the minimum treatment effect that can be deemed significant (for the given outcome), from a fixed cut-off and sample size. Such feedback will be helpful to either change the outcomes or change the population size correspondingly.

We discussed in Sections 15.1 and 15.3 that although effective imaging-derived disease markers exist (either based on machine learning models or directly computed from imaging ROIs), they may not lead to the best possible clinical trials. This is supported by the results in Table 15.9, where both rDAm and rDrm (designed to explicitly reduce the prediction variance) are compared to existing markers that have been used as trial inclusion criteria [8,6,18]. For example, ROI summaries from multiple imaging modalities have often been used as trial enrichers [7,8], and rDAm and rDrm significantly outperform these baselines (first four rows in Table 15.9). Further, [6] used SVMs to design an effective disease marker and used it as an inclusion criterion in trials. Correspondingly, we compared rDAm and rDrm to MKLm (based on a multi-kernel SVM), and the results in Table 15.9 show that baseline rDAm and rDrm as enrichers outperform MKLm, and the improvements are higher for MOCA, RAVLT, and hippocampal volume as outcomes. For our experiments, we did not adjust any of the parameters relative to the results reports earlier [4], and they were the defaults for the MKL code-base provided on the webpage (http://pages.cs.wisc.edu/~hinrichs/MKL_ADNI/) by the authors.

The necessity of incorporating multi-modal information in designing any disease markers has been reported earlier [4,2]. This is further supported by the improvement of rDAm estimates over uni-modal measures including hippocampal volume, FDG ROI summaries and florbetapir ROI summaries. These results also build upon the work of [7,8] where such unimodal imaging summaries are used for enrichment. It is possible to demonstrate that the performance gains of rDAm over [7,8] is not merely due to using three distinct modalities but also heavily influenced by the underlying machine learning architecture that exploits this information meaningfully. To see this, compare our proposed markers to the enricher "FAH" which combines three uni-modal measures, FDG, florbetapir, and hippocampal volume in Table 15.9. FAH's sample estimates are still larger than those obtained from both rDAm and rDrm, implying that the reductions are not merely due to multi-modal data or small population size, but due to the efficacy of the randomized deep learning methods introduced here, and their capacity of picking up strong decliners with high confidence and small variance.

Overall, our evaluations and the resulting trends clearly suggest that rDAm and rDrm enrichment (rDAm better among them) reduce sample sizes significantly leading to practical and cost-effective AD clinical trials. The rDA and rDr models scale

to very large dimensions, learn from only a small number of instances, and can be easily incorporated to design robust multi-modal imaging (or non-imaging, if the corresponding blocks are designed appropriately) markers. The full implementation of the framework is made available at http://pages.cs.wisc.edu/~vamsi/rda. The framework can, nevertheless, be improved further, particularly in terms of using richer pooling strategies instead of simple ridge regression, using covariate information like age, CSF levels (or FH, APOE, etc.) in the network construction itself (instead of the pre-selection setup used earlier in our experiments). An interesting extension would be to incorporate multi-modal and multi-domain (e.g., ordinal, continuous, and nominal) information directly into the rDA or rDr construction leading to multi-variate randomized deep network models. These technical issues are of independent interest and will be investigated in future works.

ACKNOWLEDGEMENTS

NIH R01 AG040396; NSF CAREER award 1252725; NIH R01 AG021155; Wisconsin Partnership Program; UW ADRC P50 AG033514; UW ICTR 1UL1RR025011. We thank the anonymous reviewers for their valuable comments.

REFERENCES

1. J. Weuve, L.E. Hebert, P.A. Scherr, D.A. Evans, Deaths in the United States among persons with Alzheimer's disease (2010–2050), Alzheimer's Dement. 10 (2) (2014) e40–e46.
2. D. Zhang, Y. Wang, L. Zhou, et al., Multimodal classification of Alzheimer's disease and mild cognitive impairment, NeuroImage 55 (3) (2011) 856–867.
3. S.J. Teipel, C. Born, M. Ewers, A.L. Bokde, M.F. Reiser, H.J. Möller, H. Hampel, Multivariate deformation-based analysis of brain atrophy to predict Alzheimer's disease in mild cognitive impairment, NeuroImage 38 (2007) 13–24.
4. C. Hinrichs, V. Singh, G. Xu, S.C. Johnson, Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population, NeuroImage 55 (2011) 574–589.
5. C. Hinrichs, N. Dowling, S. Johnson, V. Singh, MKL-based sample enrichment and customized outcomes enable smaller ad clinical trials, in: MLINI, in: Lect. Notes Comput. Sci., vol. 7263, Springer, Berlin, Heidelberg, ISBN 978-3-642-34712-2, 2012, pp. 124–131.
6. O. Kohannim, X. Hua, D.P. Hibar, S. Lee, Y.Y. Chou, A.W. Toga, C.R. Jack Jr., M.W. Weiner, P.M. Thompson, Boosting power for clinical trials using classifiers based on multiple biomarkers, Neurobiol. Aging 31 (2010) 1429–1442.
7. J.D. Grill, L. Di, P.H. Lu, C. Lee, J. Ringman, L.G. Apostolova, et al., Estimating sample sizes for predementia Alzheimer's trials based on the Alzheimer's Disease Neuroimaging Initiative, Neurobiol. Aging 34 (2013) 62–72.
8. J.D. Grill, S.E. Monsell, Choosing Alzheimer's disease prevention clinical trial populations, Neurobiol. Aging 35 (3) (2014) 466–471.

9. V. Jelic, M. Kivipelto, B. Winblad, Clinical trials in mild cognitive impairment: lessons for the future, J. Neurol. Neurosurg. Psychiatry 77 (4) (2006) 429–438.
10. R.C. Petersen, Mild cognitive impairment: current research and clinical implications, in: Seminars in Neurology, vol. 27, 2007, pp. 22–31.
11. P.S. Aisen, Clinical trial methodologies for disease-modifying therapeutic approaches, Neurobiol. Aging 32 (2011) S64–S66.
12. M.S. Albert, S.T. DeKosky, D. Dickson, B. Dubois, H.H. Feldman, N.C. Fox, A. Gamst, D.M. Holtzman, W.J. Jagust, R.C. Petersen, et al., The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the National Institute on Aging–Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease, Alzheimer's Dement. 7 (3) (2011) 270–279.
13. A.J. Mitchell, M. Shiri-Feshki, Rate of progression of mild cognitive impairment to dementia – meta-analysis of 41 robust inception cohort studies, Acta Psychiatr. Scand. 119 (4) (2009) 252–265.
14. M. Lorenzi, M. Donohue, D. Paternico, C. Scarpazza, S. Ostrowitzki, O. Blin, E. Irving, G. Frisoni, A.D.N. Initiative, et al., Enrichment through biomarkers in clinical trials of Alzheimer's drugs in patients with mild cognitive impairment, Neurobiol. Aging 31 (8) (2010) 1443–1451.
15. J.M.S. Leoutsakos, A.L. Bartlett, S.N. Forrester, C.G. Lyketsos, Simulating effects of biomarker enrichment on Alzheimer's disease prevention trials: conceptual framework and example, Alzheimer's Dement. 10 (2) (2014) 152–161.
16. N. Mattsson, U. Andreasson, S. Persson, M.C. Carrillo, S. Collins, S. Chalbot, N. Cutler, D. Dufour-Rainfray, A.M. Fagan, N.H. Heegaard, et al., CSF biomarker variability in the Alzheimer's Association quality control program, Alzheimer's Dement. 9 (3) (2013) 251–261.
17. J. Escudero, J.P. Zajicek, E. Ifeachor, Machine learning classification of MRI features of Alzheimer's disease and mild cognitive impairment subjects to reduce the sample size in clinical trials, in: Engineering in Medicine and Biology Society, EMBC, 2011 Annual International Conference of the IEEE, IEEE, 2011, pp. 7957–7960.
18. P. Yu, J. Sun, R. Wolz, D. Stephenson, J. Brewer, N.C. Fox, P.E. Cole, C.R. Jack, D.L. Hill, A.J. Schwarz, et al., Operationalizing hippocampal volume as an enrichment biomarker for amnestic mild cognitive impairment trials: effect of algorithm, test-retest variability, and cut point on trial cost, duration, and sample size, Neurobiol. Aging 35 (4) (2014) 808–818.
19. Y. Bengio, I.J. Goodfellow, A. Courville, Deep Learning, a MIT book in preparation. Draft chapters available at http://www.deeplearningbook.org/, 2015.
20. Y. Bengio, Learning deep architectures for AI, Found. Trends Mach. Learn. 2 (2009) 1–127.
21. Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
22. D. Erhan, Y. Bengio, A. Courville, P.A. Manzagol, P. Vincent, S. Bengio, Why does unsupervised pre-training help deep learning?, J. Mach. Learn. Res. 11 (2010) 625–660.
23. K. Kavukcuoglu, P. Sermanet, Y. Boureau, K. Gregor, M. Mathieu, Y. LeCun, Learning convolutional feature hierarchies for visual recognition, in: Advances in Neural Information Processing Systems, vol. 1, 2010, p. 5.
24. A. Krizhevsky, I. Sutskever, G. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems, vol. 1, 2012, p. 4.

25. G. Dahl, D. Yu, L. Deng, A. Acero, Large vocabulary continuous speech recognition with context-dependent DBN-HMMs, in: Proceedings of Acoustics, Speech and Signal Processing (ICASSP), 2011, pp. 4688–4691.
26. F. Bach, Breaking the curse of dimensionality with convex neural networks, arXiv: 1412.8690, 2014.
27. V.K. Ithapu, S. Ravi, V. Singh, On the interplay of network structure and gradient convergence in deep learning, arXiv:1511.05297, 2015.
28. H.I. Suk, D. Shen, Deep learning-based feature representation for AD/MCI classification, in: K. Mori, I. Sakuma, Y. Sato, C. Barillot, N. Navab (Eds.), MICCAI, in: Lect. Notes Comput. Sci., vol. 8150, Springer, Berlin, Heidelberg, ISBN 978-3-642-40762-8, 2013, pp. 583–590.
29. A. Gupta, M. Ayhan, A. Maida, Natural image bases to represent neuroimaging data, in: Proceedings of the 30th ICML, 2013, pp. 987–994.
30. S.M. Plis, D.R. Hjelm, R. Salakhutdinov, V.D. Calhoun, Deep learning for neuroimaging: a validation study, arXiv:1312.5847, 2013.
31. L.M. Friedman, C.D. Furberg, D.L. DeMets, Sample size, in: Fundamentals of Clinical Trials, Springer, 2010, pp. 133–167.
32. P. Vincent, H. Larochelle, I. Lajoie, Y. Bengio, P. Manzagol, Stacked denoising autoencoders: learning useful representations in a deep network with a local denoising criterion, J. Mach. Learn. Res. 11 (2010) 3371–3408.
33. N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov, Dropout: a simple way to prevent neural networks from overfitting, J. Mach. Learn. Res. 15 (1) (2014) 1929–1958.
34. T.V. Sakpal, Sample size estimation in clinical trial, Perspect. Clin. Res. 1 (2) (2010) 67–69.
35. M. Minsky, S. Papert, Perceptrons, 1969.
36. G.E. Dahl, T.N. Sainath, G.E. Hinton, Improving deep neural networks for LVCSR using rectified linear units and dropout, in: 2013 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2013, pp. 8609–8613.
37. L. Bottou, Large-scale machine learning with stochastic gradient descent, in: Proceedings of COMPSTAT'2010, Springer, 2010, pp. 177–186.
38. Y.A. LeCun, L. Bottou, G.B. Orr, K.R. Müller, Efficient BackProp, in: Neural Networks: Tricks of the Trade, Springer, 2012, pp. 9–48.
39. C.M. Bishop, Pattern Recognition and Machine Learning, Springer, 2006.
40. N. Le Roux, Y. Bengio, Representational power of restricted Boltzmann machines and deep belief networks, Neural Comput. 20 (6) (2008) 1631–1649.
41. P. Baldi, P. Sadowski, The dropout learning algorithm, Artif. Intell. 210 (2014) 78–122.
42. S. Boyd, L. Vandenberghe, Convex Optimization, Cambridge University Press, 2004.
43. S. Kay, Fundamentals of Statistical Signal Processing: Estimation Theory, Prentice Hall, 1993.
44. S.T. Smith, Statistical resolution limits and the complexified Cramér–Rao bound, IEEE Trans. Signal Process. 53 (5) (2005) 1597–1609.
45. T.G. Dietterich, Ensemble methods in machine learning, in: Multiple Classifier Systems, 2000, pp. 1–15.
46. D. Erhan, P. Manzagol, Y. Bengio, S. Bengio, P. Vincent, The difficulty of training deep architectures and the effect of unsupervised pre-training, in: Proceedings of the International Conference on Artificial Intelligence and Statistics, 2009, pp. 153–160.

47. J. Ngiam, A. Coates, A. Lahiri, B. Prochnow, Q.V. Le, A.Y. Ng, On optimization methods for deep learning, in: Proceedings of the 28th International Conference on Machine Learning (ICML-11), 2011, pp. 265–272.
48. C. Hinrichs, V. Singh, G. Xu, S.C. Johnson, A.D.N. Initiative, et al., Predictive markers for AD in a multi-modality framework: an analysis of MCI progression in the ADNI population, NeuroImage 55 (2) (2011) 574–589.
49. R. Bourgon, R. Gentleman, W. Huber, Independent filtering increases detection power for high-throughput experiments, Proc. Natl. Acad. Sci. USA 107 (21) (2010) 9546–9551.
50. C.R. Jack, D.S. Knopman, W.J. Jagust, R.C. Petersen, M.W. Weiner, P.S. Aisen, L.M. Shaw, P. Vemuri, H.J. Wiste, S.D. Weigand, et al., Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers, Lancet Neurol. 12 (2) (2013) 207–216.
51. M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, E. Liu, et al., The Alzheimer's disease neuroimaging initiative: a review of papers published since its inception, Alzheimer's Dement. 9 (5) (2013) e111–e194.
52. J. Ashburner, VBM Tutorial, 2010.
53. J. Ashburner, G. Barnes, C. Chen, J. Daunizeau, G. Flandin, K. Friston, et al., SPM8 Manual, 2008.
54. M.W. Weiner, D.P. Veitch, P.S. Aisen, L.A. Beckett, N.J. Cairns, J. Cedarbaum, R.C. Green, D. Harvey, C.R. Jack, W. Jagust, et al., 2015 update of the Alzheimer's disease neuroimaging initiative: a review of papers published since its inception, Alzheimer's Dement. 11 (6) (2015) e1–e120.
55. C. Hinrichs, V. Singh, J. Peng, S. Johnson, Q-MKL: matrix-induced regularization in multi-kernel learning with applications to neuroimaging, in: Advances in Neural Information Processing Systems, 2012, pp. 1421–1429.
56. S. Klöppel, C.M. Stonnington, C. Chu, B. Draganski, R.I. Scahill, J.D. Rohrer, N.C. Fox, C.R. Jack, J. Ashburner, R.S. Frackowiak, Automatic classification of MR scans in Alzheimer's disease, Brain 131 (3) (2008) 681–689.
57. C. Davatzikos, P. Bhatt, L.M. Shaw, K.N. Batmanghelich, J.Q. Trojanowski, Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification, Neurobiol. Aging 32 (2011) 2322.e19.
58. E.C. Edmonds, L. Delano-Wood, L.R. Clark, A.J. Jak, D.A. Nation, C.R. McDonald, D.J. Libon, R. Au, D. Galasko, D.P. Salmon, et al., Susceptibility of the conventional criteria for mild cognitive impairment to false-positive diagnostic errors, Alzheimer's Dement. 11 (4) (2015) 415–424.
59. W. Huang, C. Qiu, E. von Strauss, B. Winblad, L. Fratiglioni, APOE genotype, family history of dementia, and Alzheimer disease risk: a 6-year follow-up study, Arch. Neurol. 61 (12) (2004) 1930–1934.

NOTES

1. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a 60 million, 5-year public–private partnership. The primary goal of ADNI has been to test whether MRI, PET, other biological markers, including clinical and neuropsychological assessment can be combined to measure the progression of MCI and early AD.

2. http://adni.loni.usc.edu/methods/documents/mri-protocols/, http://adni.loni.usc.edu/methods/pet-analysis/pet-acquisition/.
3. http://adni.loni.usc.edu/methods/documents/, http://www.alz.org/research/science/earlier_alzheimers_diagnosis.asp.
4. FDG ROIs include Left Angular Lobe, Right Angular Lobe, Left Temporal Lobe, Right Temporal Lobe, and Cingulate. AV45 ROIs include Frontal Lobe, Temporal Lobe, Parietal Lobe, and Cingulate gray matter. The corresponding ROI measures are summed up to obtain single global summary for each of FDG and AV45.