MKL-Based Sample Enrichment and Customized Outcomes Enable Smaller AD Clinical Trials

Chris Hinrichs[1,2], N. Maritza Dowling[2,3],
Sterling C. Johnson[3], and Vikas Singh[2,1]

[1] Depts. of Computer Sciences
[2] Biostatistics & Med. Informatics
[3] Medicine
University of Wisconsin–Madison
hinrichs@cs.wisc.edu, scj@medicine.wisc.edu,
{mdowlin,vsingh}@biostat.wisc.edu

Abstract. Recently, the field of neuroimaging analysis has seen a large number of studies which use machine learning methods to make predictions about the progression of Alzheimer's Disease (AD) in mildly demented subjects. Among these, Multi-Kernel Learning (MKL) has emerged as a powerful tool for systematically aggregating diverse data views, and several groups have shown that MKL is uniquely suited to combining different imaging modalities into a single learned model. The next phase of this research is to employ these predictive abilities to design more efficient clinical trials. Two issues can hamper a trial's effectiveness: the presence of non-pathological subjects in a study, and the sensitivity of the chosen outcome measure to the pathology of interest. We offer two approaches for dealing with these issues: *trial enrichment*, in which MKL-derived predictions are used to screen out subjects unlikely to benefit from a treatment; and *custom outcome measures* which use an SVM to select a *weighted* voxel average for use as an outcome measure. We provide preliminary evidence that these two strategies can lead to significant reductions in sample sizes in hypothetical trials, which directly gives reduced costs and higher efficiency in the drug development cycle.

1 Introduction

There is an extensive (and growing) interest focused on developing Disease Modifying (DM) interventions for Alzheimer's Disease (AD) in the hope of slowing or deferring the associated cognitive and functional decline. As the neurological atrophy caused by AD is irreversible, early diagnosis is critical. In its early stages, AD progresses slowly – with measurable anatomical changes *preceding* cognitive losses by up to two decades – meaning that subjects enrolled in a clinical trial must be in as early a phase of AD as possible, in order to gauge the efficacy of a new pharmaceutical or other treatment when there is still time to avert cognitive decline and AD-type dementia. Thus, subjects diagnosed with Mild Cognitive Impairment (MCI), often a precursor to AD, are a primary focus in treatment

G. Langs et al. (Eds.): MLINI 2011, LNAI 7263, pp. 124–131, 2012.

procedures. Paradoxically, the emphasis on early-stage AD forces us to estimate pathological burdens when signs such as cortical atrophy or hypometabolism are at their weakest, making the need for reliable markers of disease progression a dominant concern.

Significant effort has been directed towards developing predictive markers which utilize brain imaging, (including *e.g.*, Magnetic Resonance (MR) for macrostructural information, and ^{18}Fluoro-Deoxy Glucose Positron Emission Tomography (FDG-PET) for metabolic measures) combined with machine learning algorithms. The learned model's parameters can be interpreted as a *discriminative* disease pattern, while its outputs predict which subjects will develop AD [1,2]. The machine learning approach is motivated by the observation that AD, (and other neurodegenerative disorders,) have many confounding factors, as well as significant heterogeneity. By extracting a pattern which specifically differentiates diseased and healthy populations, machine learning methods can avoid some of these pitfalls.

As methods of discriminating subjects who already have AD from controls have become more accurate, attention has shifted to the more difficult problem of discriminating MCI subjects from controls [3,4], and hardest of all, discriminating which MCI subjects will convert to AD [5,6] from the patients whose diagnoses will remain stable. One way of attacking this problem is to combine different imaging modalities for improved accuracy without incurring statistical penalties relating to the increased dimensionality of the data [7,5,6]. More recently, the focus has again shifted towards translational applications of imaging-derived predictive measures in clinical trials, with the promise of increasing sensitivity and relaxing cohort size requirements. In this context, discriminating converters from non-converters is particularly relevant because a large subgroup of non-converters can mask real treatment effects; even if the treatment is effective, it will have little or no measurable effect on subjects who do not suffer from the disease. Given that on average only 10–15% of MCI subjects convert annually, we can expect this to be a serious problem. For example, Visser et al. [8] suspected that several AD trials may have failed for exactly this reason.

Second, consider the difficulty of using cognitive markers as an outcome measure. A common practice in clinical trials is to measure changes over time in various neuropsychological status measures such as the Mini-Mental State Exam (MMSE) [9]. Unfortunately, such measures are subject to a large amount of inter- and intra-subject variation, and produce meager group effect sizes when measured by annual change. Recent results have shown [10,11] that with imaging-based outcome measures, cohort sizes can be greatly reduced – by up to a factor of 8 in [11]. We propose to move beyond these studies by using a predictive marker based on learning methods rather than summary statistics of atrophy over Regions Of Interest (ROIs). In the following we present evidence that by using better imaging-derived markers we can **enrich** the sample population to remove a large portion of the non-converters, and we can train more sensitive **custom outcome measures** in place of ROI summary statistics.

The contributions of this paper are: **(1)** We propose a new method of using a *multi-modality* predictive marker as a selection criterion for clinical trial sample enrichment; **(2)** we propose a new discriminative marker based on Tensor-Based Morphometry (TBM) to produce custom outcome measures; and **(3)** through experiments on the ADNI image dataset, we show that substantial reductions in sample sizes over standard methods are possible.

2 Preliminaries

2.1 Multi Kernel Learning (MKL)

Support Vector Machines (SVM) [12] classify subjects into separate categories (*e.g.*, diseased and healthy) by finding a separating hyperplane which balances *classification accuracy* with *separating margin* between the classes. In the AD setting, each subject's brain image is a feature vector with each dimension corresponding to a single voxel intensity value, or other imaging-derived feature. We seek a separating hyperplane which not only places the controls on one side, and the AD subjects on the other (classification accuracy), but also puts the greatest possible distance between the two groups of points, (separating margin). We can express the SVM training procedure as a quadratic program (QP) of the form,

$$\begin{aligned} \min_{\mathbf{w},\mathbf{b},\xi\geq 0} \quad & \frac{\|\mathbf{w}\|^2}{2} + C\sum_i \xi_i \\ \text{s.t.} \quad & y_i\left(\mathbf{w}^T x_i + b\right) + \xi_i \geq 1,\ \forall i \end{aligned} \tag{1}$$

where each subject / label pair is written as (x_i, y_i), $\mathbf{w}$ is a weight vector which determines the separating hyperplane, and ξ is a vector of "slack variables" which allow the algorithm to make errors in case the data are not completely separable. Each constraint encodes the desired outcome that one training subject be placed at least a "unit" distance away from the hyperplane, (with ξ_i taking up the slack). The actual "units" by which the margin is measured are given by $1/\|\mathbf{w}\|$, which we maximize by minimizing $\|\mathbf{w}\|^2/2$. C expresses a trade-off between accuracy and margin. The dual problem is:

$$\begin{aligned} \min_{0\leq\alpha\leq C} \quad & \mathbf{1}^T\alpha - (\alpha\circ\mathbf{y})^T K(\alpha\circ\mathbf{y}) \\ \text{s.t.} \quad & \mathbf{y}^T\alpha = 0 \end{aligned} \tag{2}$$

where $\circ$ denotes element-wise multiplication. Notice that the examples x_i only appear as inner products $x_i^T x_j$, which we substitute with a **kernel matrix** K. Various non-linear functions of the kernel matrix ($K \succeq 0$) correspond to non-linear transformations of the data, allowing for a richer set of classifiers.

In the context of multi-modality kernel learning, each separate modality generates a series of linear and non-linear kernels. A common methodology for combining these kernels is Multi-Kernel Learning (MKL) [5,6,13,14]. MKL solves

this problem by simultaneously choosing a linear combination of kernels, *and* estimating a max-margin classifier. To preserve the margin, the coefficients β must also be regularized, which gives the MKL primal problem [14]:

$$\min_{\mathbf{w},\xi,\beta,b} \quad \sum_m \frac{||\mathbf{w_m}||_2^2}{\beta_m} + C\sum_i \xi_i + \|\beta\|_p^2 \tag{3}$$

$$\text{s.t.} \quad y_i\left(\sum_m \mathbf{w_m}^T \phi_m(x_i) + b\right) + \xi_i \geq 1 \; \forall i \tag{4}$$

where ϕ_m transforms the observed data to the m^{th} kernel space. The constraints now express the classifier as a sum of contributions from each kernel space. Thus, we construct a series of kernels based on all imaging modalities, distinct image processing pipe-lines, feature selection methods, and kernel functions, which are then used to train an MKL classifier. We will interpret the output from the model on MCI participants as a Multi Kernel Learning based Inclusion Criterion (MKL-IC). Subjects who do not satisfy this inclusion criteria (have a high likelihood of *not* converting) will not be included in the trial – leading to an "enriched" population. Our objective then is to analyze an outcome measure of interest (*e.g.*, atrophy, cognitive decline), calculate its variance in the enriched cohort, and assess the number of subjects needed to observe a pre-selected level of difference in that outcome due to the treatment at a given power.

2.2 Power Calculation

Having selected an outcome measure for use in a clinical trial, the principal question becomes, what number of subjects (sample size) do we need to recruit in order to observe (*e.g.*, at 80% power) the induced variations in the outcome measure. This calculation is transparent to the actual drug under study, and is fully determined by the *variance* and *effect size* (difference of means between placebo / treatment groups). For a two sample t-test, the power function for testing the null hypothesis $H_0 : \delta = \mu_t - \mu_p = 0$ against the alternate hypothesis $H_a : \delta = \mu_t - \mu_p \neq 0$ is given by $P\{z > Z_{\alpha/2} \text{ or } z < -Z_{\alpha/2}\} = 1 - \Phi[Z_{\alpha/2}(\delta/\sigma)(\sqrt{n/2})] + \Phi[-Z_{\alpha/2} - (\delta/\sigma)(\sqrt{n/2})]$, where Φ is the standard normal cumulative distribution function, $z = \frac{\bar{X}_t - \bar{X}_p}{\sigma\sqrt{2/n}}$ is the test statistic, and $Z_{\alpha/2}$ is the upper $\alpha/2$ percentile from the standard normal distribution. After a simple algebraic manipulation and assuming a two-tailed test, the sample size per group is,

$$n = \frac{2\sigma^2(Z_{1-\alpha/2} + Z_{1-\beta})^2}{(\delta^*)^2}.$$

where $(1-\beta)$ is the desired power and $\delta^* = \lambda(\delta)$ is the effect size. In this context, λ denotes the expected percentage of reduction in mean annual atrophy rate (as determined by the outcome measure,) δ the estimated average annual change for the group, and σ^2 the variance of the outcome measure across individuals. Note that these calculations depend only on the mean and variance of the outcome

measure (among the selected cohort), and do not depend on the type of treatment being administered, so long as it can be assumed to affect the measured atrophy rate. Crucially, this allows us to use data in which no treatment is under trial to evaluate various outcome measures and inclusion criteria in terms of their effects on required sample sizes.

2.3 Custom TBM Measure

In addition to the MKL-derived screening criteria, we also examined the effectiveness of an SVM-derived custom outcome measure. Rather than the voxel intensity mean, (which, in the case of TBM data corresponds to a kind of average atrophy), we chose a *weighted* average of atrophy. The rationale in doing so is the observation that SVMs choose a linear function of the observed features which is far more discriminative than a simple weighted average; if this were not so, then there would be no need for SVMs in the first place. To construct this outcome measure, we trained an SVM classifier using the AD and normal control populations, and negated voxels with negative mean change so that the SVM weights came out all positive. We then normalized the SVM weights to sum to 1, so that our outcome measure can be considered a legitimate weighted average.

3 Experimental Design

Our experiments to assess the efficacy of the enrichment procedure and custom outcome measure described above were conducted on an extensive dataset of different image types and cognitive scores acquired as part of Alzheimer's Disease Neuroimaging Initiative (ADNI) [15] (ADNI is a public-private partnership to evaluate whether brain imaging can detect early signs of AD better than cognitive and biological measures [10,11,16].) Our goal was to calculate sample sizes required in a hypothetical placebo-controlled parallel clinical trial to observe a given reduction in the rate of atrophy (via an outcome measure) at a given level of statistical power. To highlight potential gains, we provide power calculations both with and without the new MKL-derived inclusion criteria, which we refer to as "MKL-IC", and custom SVM-derived outcome measure, which we refer to as "Custom-SVM". For comparison we also used mean TBM values in the chosen ROI, as in [10,11] ("Mean TBM" in Table 1.)

3.1 Dataset and Pre-processing

Our data included 48 AD cases, 66 controls and 119 MCI patients. All scans were non-linearly warped to a DARTEL template. Voxel-Based Morphometry (VBM) and Tensor-Based Morphometry (TBM) processing pipelines were applied to MR data to extract baseline Gray Matter(GM) density and longitudinal deformation maps. FDG-PET scans from baseline and at 24 months were also included, for a total of four groups of images – which provided the kernels used in our MKL

model (for the MKL-IC measure). For our evaluations, we computed t-statistics from each voxel using the AD and control population *only*, and then thresholded the voxels at $p < 0.05$ to produce a statistical ROI. TBM deformation maps were used to compute outcome measures of interest (atrophy). A natural question is whether TBM can be used both for learning the MKL-IC (albeit from AD and controls) *and* as an outcome measure for the MCI group. Making this choice is similar to the common practice of using *e.g.*, hippocampal volume measures both as covariates and atrophy as an outcome measure [16]. However, if desired (and in the interest of being more conservative), one may prefer not to use such measures in both MKL-IC and in outcome measures. We discuss both results in the following section (see "TBM/No-TBM" columns in Table 1). MKL was implemented with the Shogun package in Matlab [13], using a (non-sparse) 2-norm regularizer on kernel mixing weights. AD and control subjects were used for training the classifier, (*i.e.*, learning the disease pattern), and for feature selection (*i.e.*, for selecting Regions of Interest (ROI)). We then computed the classifier's output which provided the desired MKL Inclusion Criterion (MKL-IC). We retained the 25% of subjects whose MKL-IC was most indicative of an AD-like pattern of atrophy, and excluded the remainder. This choice is consistent with the number of subjects expected to convert to AD within 24 months.

3.2 Exploratory Analysis

Next, we provide a brief intuition on the role of the number of voxels used, as well as the trade-off between screening out subjects for greater enrichment, versus controlling the number of subjects at screening. In addition to estimating required sample sizes for fixed parameter values (25% inclusion, TBM voxels with $p < 0.05$ used in computing the outcome measures), we also computed a map of sample cohort sizes for a range of voxel selection thresholds and number of subjects included (and excluded). This analysis was exploratory in nature, (the voxels and inclusion quantile having been chosen before-hand,) and allows us to examine *qualitatively* (for this particular dataset), the choices available. The result of this analysis is shown in Fig. 1. Note the decreasing trend in n80 numbers (*i.e.*, sample sizes at 80% power – shown as cooler colors) as the inclusion criteria become more strict (*i.e.*, excluding more stable MCI subjects) highlighting the value of sample enrichment for improving detection of effects on atrophy. Notably, there appears to be a trough at about the top 25%, which is consistent with our assumptions.

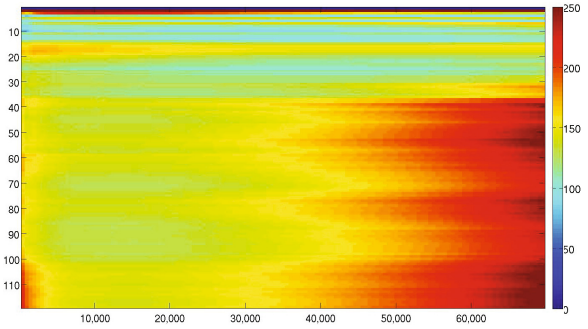

Fig. 1. Cohort sizes as a function of number of TBM voxels (x-axis), and number of MCI subjects (y-axis)

Table 1. Estimated sample cohort sizes for multi-modal inclusion criteria. TBM / NO TBM refers to whether longitudinal TBM-derived kernels were used in computing the MKL-IC. "Baseline" MKL-IC was derived *only* from data available at Month 0. Custom SVM is an SVM-derived outcome measure (weighted average over ROI).

Outcome measure	**Mean TBM**			**Custom-SVM**	**ADAS-Cog**	**MMSE**	**Schott et al. [16]**
Inclusion Criterion	**MKL-IC TBM**	**MKL-IC No TBM**	**MKL-IC Baseline**	**MKL-IC Baseline**	–	–	–
Power							
0.80	71	90	166	**88**	1,023	1,557	122
0.85	80	103	190	**100**	1,170	1,781	–
0.90	94	121	222	**117**	1,369	2,084	–

4 Results and Discussion

Table 1 presents the main results. Our primary concern is the number of subjects needed (per arm) to detect a 25% reduction in atrophy resulting from treatment. Using standard clinical and cognitive measures would require anywhere from 1000 to over 2000 subjects *per arm* to detect the treatment effect with power from 80% to 90%, type I error rate of 0.05. In contrast, by using enriched samples and imaging-based outcomes, dramatic reductions are achievable. Even without using *any* longitudinal inclusion criteria, we can reduce sample sizes by a factor of 5 to 10 (see column "Mean TBM /MKL-IC Baseline"). We also see that further improvements can result from using an SVM-derived weighted statistical ROI. Column "Custom SVM" shows numbers that are comparable to column "No TBM", (in which longitudinal FDG-PET data were used in the MKL-IC, but not longitudinal TBM data,) which strongly suggests that the gain in sensitivity from using an SVM-derived outcome measure is comparable to using longitudinal data in the inclusion criterion itself. These results compared favorably to recently reported findings [16,11,10] – our study uses *only* MCI participants, which are a more challenging group than AD participants.

In general, clinical trials designed to study the effect of DM compounds in AD are likely to require large sample sizes, and long-term duration [17]. Most DM trials estimate sample sizes using a measure of cognition as primary endpoint (see `clinicaltrials.gov`). Instead, we propose a new multi-modality screening criterion to *enrich* the sample by screening out subjects unlikely to benefit from the treatment, *and*, move beyond average ROI values to use *weighted* ROIs derived from an SVM. Our evaluations on ADNI data have shown that there exists a significant potential to improve on current practices.

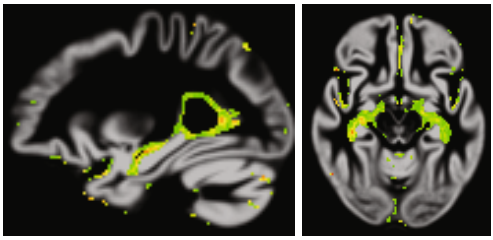

Fig. 2. Brain regions used in the TBM custom outcome measure. (Custom-SVM)

References

1. Misra, C., Fan, Y., Davatzikos, C.: Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short-term conversion to AD: Results from ADNI. Neuroimage 44(4), 1415–1422 (2008)
2. Schroeter, M.L., Stein, T., Maslowski, N., Neumann, J.: Neural correlates of alzheimer's disease and mild cognitive impairment: A systematic and quantitative meta-analysis involving 1351 patients. Neuroimage 47(4), 1196–1206 (2009)
3. Davatzikos, C., Xu, F., An, Y., Fan, Y., Resnick, S.M.: Longitudinal progression of Alzheimer's-like patterns of atrophy in normal older adults: the SPARE-AD index. Brain 132(8), 2026–2035 (2009)
4. Querbes, O., Aubry, F., Pariente, J., Lotterie, J.A., Demonet, J.F., Duret, V., Puel, M., Berry, I., Fort, J.C., Celsis, P.: Early diagnosis of alzheimer's disease using cortical thickness: impact of cognitive reserve. Brain 132(8), 2036–2047 (2009)
5. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C.: Predictive markers for AD in a multi-modality framework: An analysis of MCI progression in the ADNI population. Neuroimage 55(2), 574–589 (2011)
6. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal Classification of Alzheimer's Disease and Mild Cognitive Impairment. NeuroImage 55(3), 856–867 (2011)
7. Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., et al.: Heterogeneous data fusion for alzheimer's disease study. In: Proceeding of the 14th ACM SIGKDD, pp. 1025–1033. ACM (2008)
8. Visser, P.J., Scheltens, P., Verhey, F.R.J.: Do MCI criteria in drug trials accurately identify subjects with predementia Alzheimer's disease? Journal of Neurology, Neurosurgery & Psychiatry 76(10), 1348 (2005)
9. Petersen, R.C., Thomas, R.G., Grundman, M., et al.: Donepezil and vitamin E in the treatment of mild cognitive impairment. N. Engl. J. Med. 352, 2379–2388 (2005)
10. Kohannim, O., Hua, X., Hibar, D.P., Lee, S., Chou, Y.Y., Toga, A.W., Jack Jr., C.R., Weiner, M.W., Thompson, P.M.: Boosting power for clinical trials using classifiers based on multiple biomarkers. Neurobiology of Aging 31(8), 1429–1442 (2010)
11. Hua, X., Lee, S., Yanovsky, I., Leow, A.D., et al.: Optimizing power to track brain degeneration in Alzheimer's disease and mild cognitive impairment with TBM: An ADNI study of 515 subjects. Neuroimage 48(4), 668–681 (2009)
12. Cortes, C., Vapnik, V.: Support-vector networks. Machine Learning 20(3), 273–297 (1995)
13. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. Journal of Machine Learning Research 7, 1531–1565 (2006)
14. Kloft, M., Brefeld, U., Sonnenburg, S., Zien, A.: ℓ_p-Norm Multiple Kernel Learning. JMLR 12, 953–997 (2011)
15. Mueller, S.G., Weiner, M.W., Thal, L.J., Petersen, R.C., et al.: Ways toward an early diagnosis in Alzheimer's disease: The Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's and Dementia: The Journal of the Alzheimer's Association 1(1), 55–66 (2005)
16. Schott, J.M., Bartlett, J.W., Barnes, J., Leung, K.K., Ourselin, S., Fox, N.C.: Reduced sample sizes for atrophy outcomes in alzheimer's disease trials: baseline adjustment. Neurobiology of Aging 31(8), 1452–1462 (2010)
17. Zhang, R.Y., Leon, A.C., Chuang-Stein, C., Romano, S.J.: A new proposal for randomized start design to investigate disease-modifying therapies for Alzheimer disease. Clinical Trials 8(1), 5 (2011)