Predictive Modeling of Anatomy with Genetic and Clinical Data

Adrian V. Dalca[1], Ramesh Sridharan[1], Mert R. Sabuncu[1,2], and Polina Golland[1,⋆]

[1] Computer Science and Artificial Intelligence Lab, EECS, MIT
[2] Martinos Center for Biomedical Imaging, Harvard Medical School

Abstract. We present a semi-parametric generative model for predicting anatomy of a patient in subsequent scans following a single baseline image. Such predictive modeling promises to facilitate novel analyses in both voxel-level studies and longitudinal biomarker evaluation. We capture anatomical change through a combination of population-wide regression and a non-parametric model of the subject's health based on individual genetic and clinical indicators. In contrast to classical correlation and longitudinal analysis, we focus on predicting new observations from a single subject observation. We demonstrate prediction of follow-up anatomical scans in the ADNI cohort, and illustrate a novel analysis approach that compares a patient's scans to the predicted subject-specific healthy anatomical trajectory.

1 Introduction

We present a method for predicting anatomy based on external information, including genetic and clinical indicators. Specifically, given only a single baseline scan of a new subject in a longitudinal study, our model predicts anatomical changes and generates a subsequent image by leveraging subject-specific genetic and clinical information. Such voxel-wise prediction opens up several new areas of analysis, enabling novel investigations both at the voxel level and at the level of derivative biomarker measures. For example, voxel level differences between the true progression of a patient with dementia and their predicted healthy anatomy highlight spatial patterns of disease. We validate our method by comparing measurements of volumes of anatomical structures based on predicted images to those extracted from the acquired scans.

Our model describes the change from a single (or *baseline*) medical scan in terms of population trends and subject-specific external information. We model how anatomical appearance changes with age on average in a population, as well as deviations from the population average using a person's *health profile*. We characterize such profiles non-parametrically based on the genotype, clinical information, and the baseline image. Subject-specific change is constructed

---

⋆ A listing of ADNI investigators is available at `http://tinyurl.com/ADNI-main`. Data used via AD DREAM Challenge: `https://www.synapse.org/#!Synapse:syn2290704`

from the similarity of health profiles in the cohort, using a Gaussian process parametrized by a population health covariance. Given the predicted change, we synthesize new images through an appearance model.

Statistical population analysis is one of the central topics in medical image computing. The classical correlation-based analysis has yielded important characterization of relationships within imaging data and with independent clinical variables [2,11,12,14]. Regression models of object appearance have been previously used for atlas construction and population analysis [2,14]. These methods characterize population trends with respect to external variables, such as age or gender, and construct clinically relevant population averages. Longitudinal analyses also characterize subject-specific temporal effects, usually in terms of changes in the biomarkers of interest. Longitudinal cohorts and studies promise to provide crucial insights into aging and disease [11,12]. Mixed effects models have been shown to improve estimation of subject-specific longitudinal trends by using inter-population similarity [3,15]. While these approaches offer a powerful basis for analysis of biomarkers or images in a population, they require multiple observations for any subject, and do not aim to provide subject-specific predictions given a single observation. The parameters of the models are examined for potential scientific insight, but they are not tested for predictive power. In contrast, we define the problem of population analysis as predicting anatomical changes for individual subjects. Our generative model incorporates a population trend and uses subject-specific genetic and clinical information, along with the baseline image, to generate subsequent anatomical images. This prediction-oriented approach provides avenues for novel analysis, as illustrated by our experimental results.

2   Prediction Model

Given a dataset of patients with longitudinal data, and a single baseline image for a new patient, we predict follow-up anatomical states for the patient. We model anatomy as a phenotype y that captures the underlying structure of interest. For example, y can be a low-dimensional descriptor of the anatomy at each voxel. We assume we only have a measurement of our phenotype at baseline y_b for a new subject. Our goal is to predict the phenotype y_t at a later time t. We let x_t be the subject age at time t, and define $\Delta x_t = x_t - x_b$ and $\Delta y_t = y_t - y_b$. We model the change in phenotype y_t using linear regression:

$$\Delta y_t = \Delta x_t \beta + \epsilon, \tag{1}$$

where β is the subject-specific regression coefficient, and noise $\epsilon \sim \mathcal{N}(0, \sigma^2)$ is sampled from zero-mean Gaussian distribution with variance σ^2.

2.1   Subject-Specific Longitudinal Change

To model subject-specific effects, we define $\beta = \bar{\beta} + H(g, c, f_b)$, where $\bar{\beta}$ is a global regression coefficient shared by the entire population, and H captures a

deviation from this coefficient based on the subject's genetic variants g, clinical information c, and baseline image features f_b.

We assume that patients' genetic variants and clinical indicators affect their anatomical appearance, and that subjects with similar health profiles exhibit similar patterns of anatomical change. We let $h_G(\cdot)$, $h_C(\cdot)$, $h_I(\cdot)$ be functions that capture genetic, clinical and imaging effects on the regression coefficients:

$$H(g, c, I_b) = h_G(g) + h_C(c) + h_I(f_b). \tag{2}$$

Combining with (1), we arrive at the full model

$$\Delta y_t = \Delta x_t \bar{\beta} + \Delta x_t \big(h_G(g) + h_C(c) + h_I(f_b) \big) + \epsilon, \tag{3}$$

which captures the population trend $\bar{\beta}$, as well as the subject-specific deviations $[h_G(\cdot), h_C(\cdot), h_I(\cdot)]$.

For a longitudinal cohort of N subjects, we group all T_i observations for subject i to form $\Delta \boldsymbol{y}_i = [y_{i_1}, y_{i_2}, ... y_{i_{T_i}}]$. We then form the global vector $\Delta \boldsymbol{y} = [\Delta \boldsymbol{y}_1, \Delta \boldsymbol{y}_2, ..., \Delta \boldsymbol{y}_N]$. We similarly form vectors $\Delta \boldsymbol{x}$, $\boldsymbol{h}_G$, $\boldsymbol{h}_C$, $\boldsymbol{h}_I$, $\boldsymbol{g}$, $\boldsymbol{c}$, $\boldsymbol{f_b}$ and $\boldsymbol{\epsilon}$, to build the full regression model:

$$\Delta \boldsymbol{y} = \Delta \boldsymbol{x} \bar{\beta} + \Delta \boldsymbol{x} \odot (\boldsymbol{h}_G(\boldsymbol{g}) + \boldsymbol{h}_C(\boldsymbol{c}) + \boldsymbol{h}_I(\boldsymbol{f_b})) + \boldsymbol{\epsilon}, \tag{4}$$

where $\odot$ is the Hadamard, or element-wise product. This formulation is mathematically equivalent to a General Linear Model (GLM) [9] in terms of the health profile predictors $[\boldsymbol{h}_G, \boldsymbol{h}_C, \boldsymbol{h}_I]$.

We employ Gaussian process priors to model the health functions:

$$h_D(\cdot) \sim GP(\boldsymbol{0}, \tau_D^2 K_D(\cdot, \cdot)), \tag{5}$$

where covariance kernel function $\tau_D^2 K_D(z_i, z_j)$ captures the similarity between subjects i and j using feature vectors $\boldsymbol{z}_i$ and $\boldsymbol{z}_j$ for $D \in \{G, C, I\}$. We discuss the particular form of $K(\cdot, \cdot)$ used in the experiments later in the paper.

2.2 Learning

The Bayesian formulation in (4) and (5) can be interpreted as a linear mixed effects model (LMM) [10] or a least squares kernel machine (LSKM) regression model [5,8]. We use the LMM interpretation to learn the parameters of our model, and the LSKM interpretation to perform final phenotype predictions.

Specifically, we treat $\bar{\beta}$ as the coefficient vector of fixed effects and $\boldsymbol{h}_G, \boldsymbol{h}_C$, and $\boldsymbol{h}_I$ as independent random effects. We seek the maximum likelihood estimates of parameters $\bar{\beta}$ and $\boldsymbol{\theta} = (\tau_G^2, \tau_C^2, \tau_I^2, \sigma^2)$ by adapting standard procedures for LMMs [5,8]. As standard LMM solutions become computationally expensive for thousands of observations, we take advantage of the fact that while the entire genetic and the image phenotype data is large, the use of kernels on baseline data reduces the model size substantially. We obtain intuitive iterative updates that project the residuals at each step onto the expected rate of change in likelihood, and update $\bar{\beta}$ using the best linear unbiased predictor.

2.3 Prediction

Under the LSKM interpretation, the terms $h(\cdot)$ are estimated by minimizing a penalized squared-error loss function, which leads to the following solution [5,7,8,16]:

$$h(z_i) = \sum_{j=1}^{N} \alpha_j K(z_i, z_j) \quad \text{or} \quad h = \alpha^T K \tag{6}$$

for some vector α. Combining with the definitions of the LMM, we estimate coefficients vectors α_G, α_C and α_I from a linear system of equations that involves our estimates of $\hat{\beta}$ and θ. We can then re-write (4) as

$$\Delta y = \Delta x \bar{\beta} + \Delta x \left(\alpha_G^T K_G + \alpha_C^T K_C + \alpha_I^T K_I \right) \tag{7}$$

and predict a phenotype at time t for a new subject i:

$$y_t = y_b + \Delta x_t \left[\bar{\beta} + \sum_{j=1}^{N} \alpha_{G,j} K_G(g_i, g_j) + \alpha_{C,j} K_C(c_i, c_j) + \alpha_{I,j} K_I(f_i, f_j) \right]. \tag{8}$$

3 Model Instantiation for Anatomical Predictions

The full model (3) can be used with many reasonable phenotype definitions. Here, we describe the phenotype model we use for anatomical predictions and specify the similarity kernels of the health profile.

3.1 Anatomical Phenotype

We define a voxel-wise phenotype that enables us to predict entire anatomical images. Let Ω be the set of all spatial locations v (voxels) in an image, and $I_b = \{I_b(v)\}_{v \in \Omega}$ be the acquired baseline image. We similarly define $A = \{A(v)\}_{v \in \Omega}$, to be the population atlas template. We assume each image I is generated through a deformation field Φ_{AI}^{-1} parametrized by the corresponding displacements $\{u(v)\}_{v \in \Omega}$ from the common atlas to the subject-specific coordinate frame [14], such that $I(v) = A(v + u(v))$. We further define a follow-up image I_t as a deformation Φ_{Bt} from the baseline image I_b, which can be composed to yield an overall deformation from the atlas to the follow-up scan via $\Phi_{At}^{-1} = \Phi_{AB}^{-1} \circ \Phi_{Bt}^{-1} = \{u'(v)\}_{v \in \Omega}$:

$$I_t(v) = A(v + u'(v)). \tag{9}$$

Using displacements $u'(v)$ as the phenotype of interest in (1) captures the necessary information for predicting new images, but leads to very high dimensional descriptors. To regularize the transformation and to improve efficiency, we define a low-dimensional embedding of $u'(v)$. Specifically, we assume that the atlas provides a parcellation of the space into L anatomical labels $\mathcal{L} = \{\Psi\}_{l=1}^{L}$.

We build a low-dimensional embedding of the transformation vectors $\boldsymbol{u}(v)$ within each label using PCA. We define the relevant phenotypes $\{y_{l,c}\}$ as the coefficients associated with the first C principal components of the model that capture 95% of the variance in each label, for $l = 1 \ldots L$.

We predict the phenotypes using (8). To construct a follow-up image I_t given phenotype y_t, we first form a deformation field $\hat{\Phi}_{At}^{-1}$ by reconstruction from the estimated phenotype y_t, and use $\hat{\Phi}_{At}$ assuming an invertible transformation. Using the baseline image, we predict a subsequent image via $\Phi_{Bt} = \hat{\Phi}_{At} \circ \Phi_{AB}^{-1}$. Note that we do not directly model changes in image intensity. While population models necessitate capturing such changes, we predict changes from a baseline image. We also assume that affine transformations are not part of the deformations of interest, and thus all images are affinely registered to the atlas.

3.2   Health Similarities

To fully define the health similarity term $H(\cdot, \cdot, \cdot)$, we need to specify the forms of the kernel functions $K_G(\cdot, \cdot)$, $K_C(\cdot, \cdot)$, and $K_I(\cdot, \cdot)$.

For genetic data, we employ the identical by state (IBS) kernel often used in genetic analysis [13]. Given a vector of genetic variants g of length S, each genetic locus is encoded as $g(s) \in \{0, 1, 2\}$, and

$$K_G(g_i, g_j) = \frac{1}{2S} \sum_{s=1}^{S} (2 - |g_i(s) - g_j(s)|). \tag{10}$$

To capture similarity of clinical indicators c, we form the kernel function

$$K_C(c_i, c_j) = \exp\left(-\frac{1}{\sigma_C^2}(c_i - c_j)^T \boldsymbol{W}(c_i - c_j)\right), \tag{11}$$

where diagonal weight matrix $\boldsymbol{W}$ captures the effect size of each clinical indicator on the phenotype, and σ_C^2 is the variance of the clinical factors.

We define the image feature vectors f_b as the set of all PCA coefficients defined above for the baseline image. We define the image kernel matrix as

$$K_I(f_{b,i}, f_{b,j}) = \exp\left(-\frac{1}{\sigma_I^2}\|f_{b,i} - f_{b,j}\|_2^2\right), \tag{12}$$

where σ_I^2 is the variance of the image features.

4   Experiments

We illustrate our approach by predicting image-based phenotypes based on genetic, clinical and imaging data in the ADNI longitudinal study [6] that includes two to ten follow-up scans acquired $0.5 - 7$ years after the baseline scan. We use affine registration to align all subjects to a template constructed from 145 randomly chosen subjects, and compute non-linear registration warps Φ_{AI} for
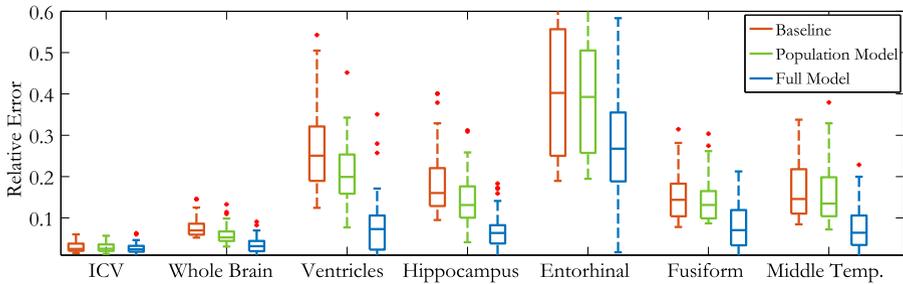
Fig. 1. Relative error (lower is better) of volume prediction for seven structures for subjects in the top decile of volume change. We report relative change between the baseline and the follow-up measurement (red), relative error in prediction using a population model (green), and the complete model (blue).

each image using ANTs [1]. We utilize a list of 21 genetic loci associated with Alzheimer's disease (AD) as the genetic vector g, and the standard clinical factors including age, gender, marital status, education, disease diagnostic, and cognitive tests, as the clinical indicator vector c. We learn the model parameters from 341 randomly chosen subjects and predict follow-up volumes on a separate set of 100 subjects. To evaluate the advantages of the proposed predictive model, we compare its performance to a population-wide linear regression model that ignores the subject-specific health profiles (i.e., $H = 0$).

4.1 Volumetric Predictions

In the first simplified experiment, we define phenotype y to be a vector of several scalar volume measurements obtained using FreeSurfer [4]. In addition to the population-wide linear regression model, we include a simple approach of using the baseline volume measurements as a predictor of the phenotype trajectory, effectively assuming no volume change with time. Since in many subjects, the volume differences are small, all three methods perform comparably when evaluated on the whole test set. To evaluate the differences between the methods, we focus on the subset of subjects with substantial volume changes, reported in Fig. 1. Our method consistently achieves smaller relative errors than the two baseline approaches.

4.2 Anatomical Prediction

We also evaluate the model for full anatomical scan prediction. To quantify prediction accuracy, we propagate segmentation labels of relevant anatomical structures from the baseline scan to the predicted scan using the predicted warps. We compare the predicted segmentation label maps with the actual segmentations of the follow-up scans. The warps computed based on the actual follow-up scans through the atlas provide an indication of the best accuracy the predictive
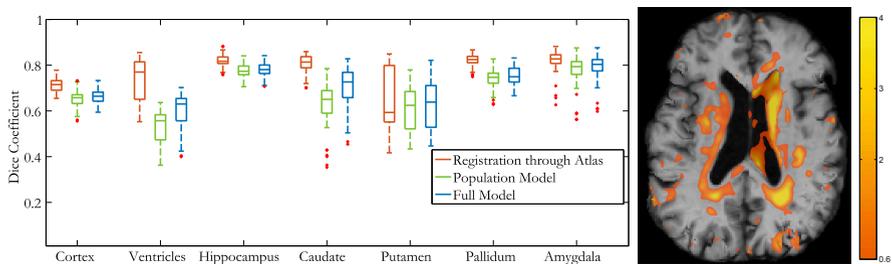
Fig. 2. Prediction results. Left: Dice scores of labels propagated through three methods for several structures implicated in AD in subjects with the most volume change for each structure. We report the prediction based on the registration of the actual follow-up scan to the atlas as an upper bound for warp-based prediction accuracy (red), predictions based on the population-wide linear regression model (green), and the full model (blue). Right: A predicted anatomical image for a patient diagnosed with AD using a healthy model. The color overlay shows the squared magnitude of the difference in predicted versus observed deformations, indicating a significantly different expansion trajectory of the ventricles.

model could achieve when using warps to represent images. Similar to the volumetric predictions, the full model offers modest improvements when evaluated on the entire test set, and substantial improvements in segmentation accuracy when evaluated in the subjects who exhibit large volume changes between the baseline scan and the follow-up scan, as reported in Fig. 2. In both experiments, all components h_g, h_c and h_I contributed significantly to the improved predictions.

Our experimental results suggest that the anatomical model depends on registration accuracy. In particular, we observe that directly registering the follow-up scan to the baseline scan leads to better alignment of segmentation labels than when transferring the labels through a composition of the transformations from the scans to the atlas space. This suggests that a different choice of appearance model may improve prediction accuracy, a promising direction for future work.

To demonstrate the potential of the anatomical prediction, we predict the follow-up scan of a patient diagnosed with dementia as if the patient were healthy. Specifically, we train our model using healthy subjects, and predict follow-up scans for AD patients. In Fig. 2 we illustrate an example result, comparing the areas of brain anatomy that differ from the observed follow-up in the predicted *healthy* brain of this AD patient. Our prediction indicates that ventricle expansion would be different if this patient had a healthy trajectory.

5   Conclusions

We present a model to predict the anatomy in patient follow-up images given just a baseline image using population trends and subject-specific genetic and clinical information. We validate our prediction method on scalar volumes and anatomical images, and show that it can be used as a powerful tool to illustrate

how a subject-specific brain might differ if it were healthy. Through this and other new applications, our prediction method presents a novel opportunity for the study of disease and anatomical development.

References

1. Avants, B.B., Tustison, N.J., Song, G., Cook, P.A., Klein, A., Gee, J.C.: A reproducible evaluation of ants similarity metric performance in brain image registration. Neuroimage 54(3), 2033–2044 (2011)
2. Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.: Population shape regression from random design data. Int. J. Comp. Vis. 90(2), 255–266 (2010)
3. Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N.: Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. International Journal of Computer Vision 103(1), 22–59 (2013)
4. Fischl, B.: Freesurfer. Neuroimage 62(2), 774–781 (2012)
5. Ge, T., Nichols, T.E., Ghosh, D., Mormino, E.C., Smoller, J.W., Sabuncu, M.R., et al.: A kernel machine method for detecting effects of interaction between multi-dimensional variable sets: An imaging genetics application. NeuroImage (2015)
6. Jack, C.R., Bernstein, M.A., Fox, N.C., Thompson, P., Alexander, G., Harvey, D., Borowski, B., Britson, P.J., L Whitwell, J., Ward, C., et al.: The alzheimer's disease neuroimaging initiative (adni): Mri methods. Journal of Magnetic Resonance Imaging 27(4), 685–691 (2008)
7. Kimeldorf, G., Wahba, G.: Some results on tchebycheffian spline functions. Journal of Mathematical Analysis and Applications 33(1), 82–95 (1971)
8. Liu, D., Lin, X., Ghosh, D.: Semiparametric regression of multidimensional genetic pathway data: Least-squares kernel machines and linear mixed models. Biometrics 63(4), 1079–1088 (2007)
9. McCullagh, P.: Generalized linear models. European Journal of Operational Research 16(3), 285–292 (1984)
10. McCulloch, C.E., Neuhaus, J.M.: Generalized linear mixed models. Wiley Online Library (2001)
11. 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 (2009)
12. Pfefferbaum, A., Rohlfing, T., Rosenbloom, M.J., Chu, W., Colrain, I.M., Sullivan, E.V.: Variation in longitudinal trajectories of regional brain volumes of healthy men and women (ages 10 to 85years) measured with atlas-based parcellation of mri. Neuroimage 65, 176–193 (2013)
13. Queller, D.C., Goodnight, K.F.: Estimating relatedness using genetic markers. Evolution, 258–275 (1989)
14. Rohlfing, T., Sullivan, E.V., Pfefferbaum, A.: Regression models of atlas appearance. In: Prince, J.L., Pham, D.L., Myers, K.J. (eds.) IPMI 2009. LNCS, vol. 5636, pp. 151–162. Springer, Heidelberg (2009)
15. Sadeghi, N., Prastawa, M., Fletcher, P.T., Vachet, C., Wang, B.: et al.: Multivariate modeling of longitudinal mri in early brain development with confidence measures. In: 2013 IEEE Inter. Symp. Biomed. Imag., pp. 1400–1403. IEEE (2013)
16. Wahba, G.: Spline models for observational data, vol. 59. SIAM (1990)