Beyond Doctors: Future Health Prediction from Multimedia and Multimodal Observations

Liqiang Nie[†], Luming Zhang[‡], Yi Yang[§], Meng Wang[‡], Richang Hong[‡] and Tat-Seng Chua[†]
[†] School of Computing, National University of Singapore, Singapore
[‡] Department of Electric Engineering and Information System, Hefei University of Technology, China
[§] Centre for Quantum Computation and Intelligent Systems (QCIS), University of Technology Sydney, Australia
{nieliqiang, zglumg, yee.i.yang, eric.mengwang, hongrc.hfut}@gmail.com, dcscts@nus.edu.sg

ABSTRACT

Although chronic diseases cannot be cured, they can be effectively controlled as long as we understand their progressions based on the current observational health records, which is often in the form of multimedia data. A large and growing body of literature has investigated the disease progression problem. However, far too little attention to date has been paid to jointly consider the following three observations of the chronic disease progression: 1) the health statuses at different time points are chronologically similar; 2) the future health statuses of each patient can be comprehensively revealed from the current multimedia and multimodal observations, such as visual scans, digital measurements and textual medical histories; and 3) the discriminative capabilities of different modalities vary significantly in accordance to specific diseases. In the light of these, we propose an adaptive multimodal multi-task learning model to co-regularize the modality agreement, temporal progression and discriminative capabilities of different modalities. We theoretically show that our proposed model is a linear system. Before training our model, we address the data missing problem via the matrix factorization approach. Extensive evaluations on a real-world Alzheimer's disease dataset well verify our proposed model. It should be noted that our model is also applicable to other chronic diseases.

Categories and Subject Descriptors

J.3 [**Life and Medical Sciences**]: Health, Medical Information Systems

General Terms

Algorithms, Performance, Experimentation

Keywords

Chronic Diseases; Multimodal Analysis; Disease Progression; Adaptive Multimodal Multi-Task Learning

MM'15, October 26–30, 2015, Brisbane, Australia.

DOI: http://dx.doi.org/10.1145/2733373.2806217.

1. INTRODUCTION

Many health conditions can be classified under the broad heading of chronic diseases, such as asthma, depression, stroke and Alzheimer's Disease (AD). They are mostly characterised by complex casuality, long latency periods, a prolonged course of illness and functional impairment. Besides financial burden, chronic diseases have profound effects on sufferers' physical and mental well-being, which often makes them difficult to carry on with daily routines[1]. Even though chronic disease does not often resolve spontaneously and is rarely cured completely[2], they progress over a long period of time to become fully established. Hence it offers us a chance for progression prediction, which is the key to moving from episodic and reactive care to a planned and proactive management [1, 14].

Disease progression prediction aims to estimate the disease statuses at some future time points, given the current multimodal observations of the patients. The severity of disease status can be measured by some clinical scores, which are designed by professionals and used as essential criteria for clinical diagnosis. Proactive care of chronic diseases would benefit from the prediction of disease progression in terms of these clinical measures. However, progression prediction in the real world faces the following challenges,

1. **Multimedia and Multimodal Observations.** Recent advances in medical technologies have made it possible to collect multimedia and multimodal data to describe the same patients [7]. For example, an echocardiogram video is acquired to provide dynamic information about the function of the heart; pictures made from Magnetic resonance imaging (MRI) scans visually show the structure problems inside the heart organs; while textual complains collected from the patients signal their symptoms. These multimedia observations, which are often multimodal data, collectively reveal the health conditions from different perspectives. Figure 1 shows a typical example of multimedia and multimodal observations for the same patient. How to effectively uncover the information embedded in the multimodal data remains a largely unaddressed research problem.
2. **Modality Confidences.** The discriminative capability of each modality differs from one to another and is heavily dependent on the type of disease. For instance, in the progression of cancer, the modality of

[1] http://www.forahealthieramerica.com/ds/impact-of-chronic-disease.html
[2] http://www.aihw.gov.au/WorkArea/DownloadAsset.aspx?id=10737421546

positron emission tomography (PET) scan, as compared to others, plays a more important role to spot a cancer, identify the stage of the cancer, show whether it has spread and help doctors make treatment plans. On the other hand, the modality of cerebrospinal fluid (CSF) measurement is more critical to multiple sclerosis.

3. **Task Relatedness.** Assume that each task concerns the prediction of a health status at one time point. Multiple tasks along a sequential timeline are not independent. Identifying and modelling their intrinsic relatedness are of vital importance, which enables the sharing of training instances among the tasks.

In addition to these three challenges, medical observations of a patient are missing at times, which could be caused by dropout or partial absence in a longitudinal study, and other private reasons. The problem of missing data may sometimes affect the prediction profoundly, and should thus be carefully addressed.

The research in disease progression modelling has been active in recent years, which can generally be divided into three categories. One is mono-modal mono-task learning. In this context, disease statuses at different time points are estimated separately by exploring a single modality [22, 10]. Neither the correlation among tasks nor the collective information across modalities is explored. Another line of efforts is the multiple task learning [35, 31]. They formulate the prediction of health statuses at a sequence of time points as a multi-task regression problem. Multi-task learning improves the learning performance by somewhat sharing the training instances among related tasks. The relatedness is modeled by assuming that they either share a common representation space or share some parameters. However, they disregard the relatedness among different modalities of a single task. The third category of approaches is the multimodal analysis [28, 30, 25]. They work towards discovering comprehensive information from several clinical modalities of the same patients for disease status prediction at one time point. Existing multimodal analysis, however, ignores the label information from other related tasks.

As an improved work, we propose a novel regression model, adaptive multimodal multi-task learning ($\mathbf{aM^2L}$), to predict chronic disease progression. Our model takes three types of prior knowledge into consideration. The first is modality agreement. Specifically, the disagreement among multimodalities are penalized, since they are supposed to reflect the same disease status. The second is adaptive modality weighting. Namely, different modalities should have different weights, which are disease specific. The last is temporal progression. In particular, the disease status progresses smoothly, and sudden changes between nearby time points should be penalized. Our model differs from [32, 11], where multi-view learning mainly focuses on semi-supervised configurations and uniformly regularizes the task relatedness instead of temporal prior. In this work, we focus on multimodal analysis in the supervised settings and we do not assume that there are abundant unlabeled data available. Before training our model, we utilize the matrix factorization approach to complete the missing data.

The contributions of this work are threefold: 1) we propose an adaptive multimodal multi-task learning approach to model the chronic disease progression from multimedia observations. The proposed method simultaneously considers modality agreement, adaptive modality weights and the temporal progression; 2) we theoretically demonstrate that our proposed model is a linear model and practically analyze its computational complexity; and 3) we validate our model on a real-world AD dataset. Our model is also applicable to other chronic diseases.

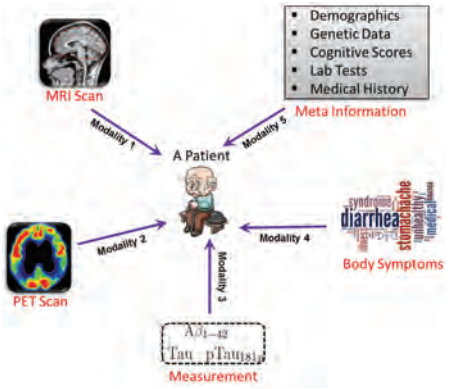

Figure 1: Illustration of multimedia and multimodal descriptions for the same patient. The modalities can be in various forms, such as visual images, digital measurements, textual descriptions, and even video records. Heterogeneous modalities reflect the health status of the same patient from diverse perspectives.

The remainder of the paper is structured as follows. Section 2 reviews the related work. Sections 3 and 4 respectively detail our data collection and the proposed disease progression model. Experimental settings and results are reported in Section 5, followed by conclusion and future work in Section 6.

2. RELATED WORK

2.1 Disease Progression modelling

Computational wellness has attracted increasing research attentions from computer science communities [19, 21, 17, 16]. Progression modeling of chronic diseases is part of the computational wellness research and it aims to predict the future disease statuses, which can be measured by the clinically defined categories [2, 5] or the continuous clinical scores [23, 6]. Broadly speaking, the existing efforts on disease progression modelling can be grouped into three categories: mono-modal mono-task learning, multi-task learning, and multimodal analysis.

Mono-modal mono-task learning in the past decade dominated the community of disease progression modelling. They estimated the disease status separately on a single modality of data, such as exploring the MRI scans to infer the targets at a future time point [23, 6]. Besides regression models, survival models were introduced in [22] to predict the future disease status of liver transplant patients by considering historical clinical variables individually. In addition, some approach [10] considered a small number of input features, and each feature was individually added to the model to examine its effectiveness. However, when the number of features is large and significant correlations among features exist, these approaches are usually suboptimal. Meanwhile, they neither considered the intrinsic correlation among tasks, nor utilized the complementary information embedded in different modalities. Their performance is thus far from satisfactory to be clinically useful.

Multi-task learning was recently proposed to model the disease progression, which utilizes the correlations among different tasks and simultaneously learns a problem together with other related problems. This often leads to a better model than learning each task separately. The key issue in multi-task learning is how to characterize and use the relatedness among multiple tasks. Two kinds of relatedness have been studied in the disease progression modelling. One is that the multiple tasks are assumed to share parameters or prior distributions of the hyper parameters. For example, Zhou et al. [35] formulated the prediction of clinical scores at a sequence of time points as a multi-task regression problem and captured the intrinsic relatedness among different tasks by a temporal group lasso regularizer. The other way of modelling the inter-task relatedness is to assume that they share a common underlying representation. For instance, the work in [31] constrained the models to share a common set of features. The work in [34] proposed a novel sparse group lasso formulation to select task-sharing and task-specific features in parallel.

Multimodal analysis was proposed to integrate clinical data from multiple channels, such as genetic, imaging and medical history. Ye et al. [28] proposed a multiple kernel learning method for integrating imaging and non-imaging data for AD study and extended the kernel framework to selecting features from heterogeneous modalities. Experiments showed that the integration of multimedia data leads to a considerable improvement in prediction accuracy. One common problem that hampers the use of multimodal analysis is the problem of missing data. To address this issue, two novel methods were introduced in [30] for joint analysis of incomplete multimodal neuroimaging data, where patients with missing measures were also kept for training. One year later, Xiang et al. [25] presented a "bi-level" learning model for multimodal block-wise missing data. It is capable of handling both feature-level and modality-level analysis.

2.2 Multi-View Multi-Task Learning

Disease progression modelling exhibits dual heterogeneities. In particular, a single learning task could have features from multiple modalities and multiple learning tasks could be related to each other through some shared components. Existing multi-task learning or multimodal analysis only captures one of the dual heterogeneities. Our proposed $\mathbf{aM^2L}$ approach well-addressed both by jointly regularizing task and modality relatedness. In fact, our model falls into the community of multi-view multi-task learning.

There are relatively sparse literatures so far on multi-task problem with multi-view data [31]. For example, He and Lawrence [9] proposed a graph-based iterative framework for multi-view multi-task learning ($IteM^2$) and applied it to text classification. $IteM^2$ projects task pairs to a new Reproducing Kernel Hilbert Space based on the common views shared by them. However, it can only deal with problems with only non-negative feature values. In addition, it is a transductive model. Hence it is unable to generate predictive models for independent and unseen testing samples. To address the intrinsic limitations of transductive models, an inductive multi-view multi-task learning model regMVMT was introduced in [32]. regMVMT regulates the view consistency on the unlabeled samples. Additional regularization functions are utilized to ensure that the learned functions are similar to each other across multiple tasks. However, without prior knowledge, simply restricting all the tasks to be similar is often inappropriate. As an extension of regMVMT, an inductive convex shared structure learning algorithm for multi-view multi-task problem (CSL-MTMV) was developed in [11]. Compared to regMVMT, CSL-MTMV considers the shared predictive structure among multiple tasks.

However, none of the methods mentioned above have been successfully applied to disease progression modelling except the one in [31]. This may be due to the facts that: 1) $IteM^2$, regMVMT and CSL-MTMV are all binary classification models, for which the extension to multi-class or regression problem is nontrivial, especially when the number of classes is large; and 2) the tasks in disease progression prediction are temporally related rather than uniformly or structurally related. Uniform and structural relations are considered in the three methods. The work in [31] treated the estimation of multiple clinical scores as different tasks and adopted the multi-task learning model [27] to learn a common feature subset. They then utilized a kernel-based multimodal data fusion method to fuse features from individual modality. They finally trained a support vector regression model to predict multiple clinical scores. However, this work does not consider the chronological progression of diseases and the separate tri-stages are hardly to reinforce each other. As a complement, we propose a unified multimodal multi-task learning framework for disease progression modelling, which regularizes task relatedness, modality agreement, and adaptive modality weights at the same time.

3. DATA COLLECTION

In this work, we selected one chronic disease, AD, due to following reasons: 1) **Prevalence**. It is the most common type of dementia, accounting for 60-80% of age-related dementia cases. It affects about 5.2 million people in the US, with a significant increase predicted in the near future if no disease-altering therapeutics are developed[3]. 2) **Mystery**. The cause of AD is poorly understood to date, especially the discriminant modality of AD. 3) **Accessibility**. The representative study data of AD can be requested from the Alzheimer's Disease Neuroimaging Initiative[4] (ADNI). ADNI is an ongoing and longitudinal study designed to develop clinical, imaging and genetic biomarkers for the early detection and tracking of AD.

To verify our model, we officially requested all the data of ADNI-1. ADNI-1 began in *2004* and its 822 participants were recruited from 59 sites across U.S. and Canada. The gender, diagnostic, ethnic and racial distributions of these enrolled participants are summarized in Table 1. We can see that the white patients dominate the participants. In addition, the distributions over age and years of education are respectively illustrated in Figure 2. It can be seen that the majorities range from 70 to 80 years old, and were well-educated before. In our study, the date recorded when the patient performed the screening in the hospital for the first time is called baseline, and the time point for the follow-up visits is denoted by the duration starting from the baseline. Taking the notation "M06" as an example, it denotes the time point six months after the first visit.

[3] https://www.alz.org/downloads/Facts_Figures_2014.pdf
[4] http://adni.loni.usc.edu/
[5] http://surfer.nmr.mgh.harvard.edu/
[6] http://www.fil.ion.ucl.ac.uk/spm/
[7] http://www.biomarkersconsortium.org/projects.php

Table 1: Demographic statistics of the 822 ADNI-1 participants, including distribution of gender, diagnostic, ethnic and racial categories. "Unkn" refers to "unknown".

Gender		Diagnostic			Ethnic Categories			Racial Categories					
Male	Female	NL	MCI	AD	Hispanic or Latino	Not Hispanic or Latino	Unkn	American Indian or Alaskan Native	Asian	Back or African American	White	More than one	Unkn
478	344	229	405	188	21	792	9	1	14	39	764	3	1

Table 2: Statistics of modalities, number of patients, and extracted features at the baseline time.

Modality	Number of patients	Feature Dimension	Feature Extractor
MRI	818	305	Image Analysis Suite FreeSurfer[5]
PET	419	116	SPM Software for visual scan[6]
CSF	415	5	ADNI Biomarker Core Lab [24]
PROT	566	147	Biomarkers Consortium Project[7]
META	818	51	Combination of Four Types[8]

The patients in ADNI-1 at their enrolment time can be in one of three diagnostics: normal elder (NL), mild cognitive impairment (MCI), and AD. NL and MCI could progress to AD. In this work, we selected 818 patients (229 NL, 401 MCI and 188 AD). Meanwhile, we utilized five modalities to collectively describe each patient, including 1.5T MRI scans (MRI), PET, CSF, proteomics (PROT) and META. Statistics over modalities, number of patients, and extracted features at the baseline time are shown in Table 2. It can be seen that each of our selected patients has at least two modalities at the same time: MRI and META. 624-D visual and textual features in total were extracted for those patients with all the five modalities available. We will detail how to complete the missing data in the experiments.

We aim to harvest their modalities at the baseline time of the given patients to predict their future health statuses in terms of MMSE and ADAS-cog scores. MMSE and ADAS-Cog are respectively the short names of Mini Mental State Examination and Alzheimer's Disease Assessment Scale cognitive subscale. They have been shown to be correlated with the underlying AD pathology and progressive deterioration of functional ability [12]. MMSE is a sensitive, valid and reliable 30-point questionnaire that is commonly used to estimate the severity and progression of cognitive impairment of AD. The ADAS-Cog is scored from 0 to 70. A larger MMSE and a smaller ADAS-Cog score usually indicates a better health status. Table 3 summarizes the number of patients available for MMSE and ADAS-Cog at different time points in our dataset. These cognitive scores are used as groundtruth to verify our model. We do not predict the health statuses at the time point M18 and M48 because of the severe data loss. Figure 3 illustrates the average evolution of health status. For ADAS-Cog cognitive progression, the AD curve grows up very fast while the MCI curve slightly rises within four years. When it comes to the NL curve, it is relatively stable. A similar pattern can be observed for the MMSE.

[8] The META modality is a fusion of 3-D demographic(age, years of education, gender), 7-D genetic (ApoE-ε4, etc.), 23-D baseline cognitive scores (CDR, FAQ, GDS, etc.), 18-D lab tests(RCT1, RCT11, RCT12, etc.)

Table 3: The number of AD, MCI, and NL patients under MMSE/ADAS-Cog at different time points.

Type	baseline	M06	M12	M18	M24	M36	M48
AD	188	176	158	0	139	11	2
MCI	401	384	362	328	304	252	33
NL	229	221	212	0	203	187	56
Total	818	781	732	328	646	450	111

4. ADAPTIVE MULTIMODAL MULTI-TASK LEARNING

In this section, we first detail the formulation of our proposed $\mathbf{aM^2L}$ approach and its optimization. We then demonstrate its linear property followed by the analysis of its computational complexity.

Before stepping into the model, we first define the notations that will be used. Assume that we have N training patients at the baseline time $\mathbf{X} = [\mathbf{x_1}, \mathbf{x_2}, ..., \mathbf{x_N}]^T$, and their corresponding disease statuses of the following M time points $\mathbf{Y} = [\mathbf{y^1}, \mathbf{y^2}, ..., \mathbf{y^M}] \in \mathbb{R}^{N \times M}$. Each patient is characterized by S collective modalities. For example, the i-th patient can be represented by $\mathbf{x_i} = [\mathbf{x_{i1}^T}, \mathbf{x_{i2}^T}, ..., \mathbf{x_{iS}^T}]^T$, where $\mathbf{x_{is}} \in \mathbb{R}^{D_s}$, and D_s denotes the feature dimension of the s-th modality. All the training patients on the s-th modality and on all the modalities are respectively represented as $\mathbf{X_s} = [\mathbf{x_{1s}}, \mathbf{x_{2s}}, ..., \mathbf{x_{Ns}}]^T \in \mathbb{R}^{N \times D_s}$ and $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_S}] \in \mathbb{R}^{N \times D}$, where $D = \sum_{s=1}^{S} D_s$. The involved notations of this work are summarized in Table 4. Our objective is to design and learn a disease progression model, which is capable of predicting the future disease statues of new incoming patients given their multiple health descriptions at the baseline. In future, we can also incorporate the followup observations to enhance the prediction model. For example, we can merge the observations at baseline, M06 and M12 as input to predict the health status at M24. However, integrating the observational data at various time points are much more challenging. This is because the data is changing and simple fusion is unable to capture the data evolution.

4.1 Problem Formulation

In a vector-wise form, we linearly define the predictive function for all patients at the time point t with the knowledge from modality s as,

$$\mathbf{f_s^t}(\mathbf{X_s}) = \mathbf{X_s}\mathbf{w_s^t}, \tag{1}$$

where $\mathbf{w_s^t} \in \mathbb{R}^{D_s}$ is the parameter vector we aim to learn. In fact, $\mathbf{w_s^t}$ explicitly reveals the weight of features extracted from modality s. To learn $\mathbf{w_s^t}$, we jointly consider the following three kinds of prior knowledge:

1. **Modality Agreement.** Heterogeneous modalities of the same patients describe their health from various perspectives, but they coherently express the same disease status. In particular, for a given patient at each

Table 4: Summary of key symbols and notations used in this work.

Notations	Descriptions
N, S, M	number of training patients, modalities and tasks/time points, respectively.
D_s, $D = \sum_{s=1}^{S} D_s$	feature dimension on the s-th modality and on all the modalities, respectively.
$\mathbf{X_s} \in \mathbb{R}^{N \times D_s}$, $\mathbf{X} = [\mathbf{X_1}, \mathbf{X_2}, ..., \mathbf{X_S}] \in \mathbb{R}^{N \times D}$	N training patients on the s-th modality and on all the modalities, respectively.
$\mathbf{y^t} \in \mathbb{R}^N$, $\mathbf{Y} = [\mathbf{y^1}, \mathbf{y^2}, ..., \mathbf{y^M}] \in \mathbb{R}^{N \times M}$	disease statuses of all the training samples on the t-th and on the all time points, respectively.
$f_s^t(\mathbf{x_{ns}})$, $\mathbf{f_s^t}(\mathbf{X_s}) \in \mathbb{R}^N$	the predictive function for patient n at time point t with the knowledge from the modality s, and its vector-wise form for all patients.
$\mathbf{w_s^t} \in \mathbb{R}^{D_s}$, $\mathbf{W_s} = [\mathbf{w_s^1}, \mathbf{w_s^2}, ..., \mathbf{w_s^M}] \in \mathbb{R}^{D_s \times M}$	the parameter vector for modality s at time point t, and the parameter matrix for modality s at all time points. They are what we aim to learn.
α_s, $\boldsymbol{\alpha} \in \mathbb{R}^S$	the weight of the s-th modality and its vector-wise form for all the modalities.
Φ, L, C, T, Ω, Ψ	overall objective function, loss function, modality agreement regularizer, temporal progression regularizer, regularization on feature weight and modality weight.
θ_1, θ_2, θ_3, θ_4	the corresponding parameters involved for C, T, Ω, and Ψ in our model.

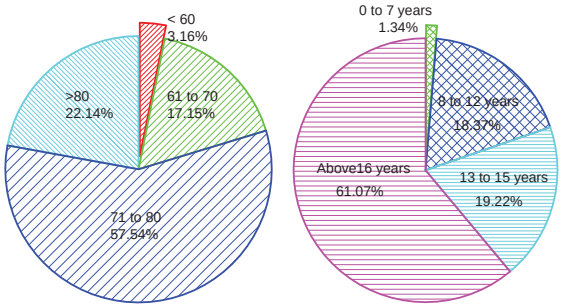

Figure 2: Distributions over age and education years of the 822 patients at the baseline time. Majorities of the enrolled patients are senile and well-educated.

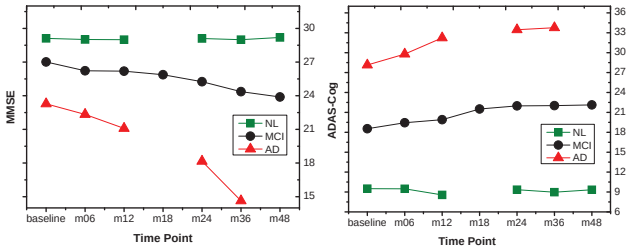

Figure 3: Cognitive score progression along time in terms of MMSE and ADAS-Cog. It is noted that some time points do not have cognitive scores.

specific time point t, the disease status estimated via exploring different modalities should be consistent.

2. **Adaptive Modality Weights.** The discriminative capabilities of each modality is disease specific. Instead of manually fixing the modality weights, which requires painful tuning, we incorporate the weighting parameters into our model as an variable and adaptively learn them via a statistical approach. Mathematically, the disease statuses for all patients at time point t are modeled by weighted fusion of the prediction results from all modalities,

$$\mathbf{f^t}(\mathbf{X}) = \sum_{s=1}^{S} \alpha_s \mathbf{f_s^t}(\mathbf{X_s}) = \sum_{s=1}^{S} \alpha_s \mathbf{X_s} \mathbf{w_s^t}, \quad (2)$$
$$\text{s.t.} \quad \sum_{s=1}^{S} \alpha_s = 1,$$

where α_s is the weight for the s-th modality of the given disease. We intentionally do not constrain α_s to be larger than zero, because in this way we can infer which modality is negatively correlated with the given chronic disease.

3. **Temporal Progression.** Chronic disease is of long duration and generally in smooth progression[9]. Hence the sudden changes of disease statuses between neighbouring time points should be penalized.

To unify the aforementioned prior knowledge, we formulate our objective function $\Phi(\mathbf{w_s^t}, \boldsymbol{\alpha})$ as,

$$L(\mathbf{w_s^t}, \boldsymbol{\alpha}) + \theta_1 C(\mathbf{w_s^t}) + \theta_2 T(\mathbf{w_s^t}) + \theta_3 \Omega(\mathbf{w_s^t}) + \theta_4 \Psi(\boldsymbol{\alpha}). \quad (3)$$

[9] http://www.hpb.gov.sg/HOPPortal/health-article/3396

The first term is a squared loss function that measures the empirical error on the training patients. As reported in [26], the squared loss usually yields good performance as other complex loss functions. We thus adopt the squared loss in our algorithm for simplicity and efficiency,

$$L(\mathbf{w_s^t}, \boldsymbol{\alpha}) = \frac{1}{2} \sum_{t=1}^{M} \| \mathbf{y^t} - \sum_{s=1}^{S} \alpha_s \mathbf{X_s} \mathbf{w_s^t} \|^2 . \quad (4)$$

In Eqn.(3), the second and third term respectively controls the consistency among various modalities and progressive changes between adjacent time points,

$$\begin{cases} C(\mathbf{w_s^t}) = \frac{1}{2} \sum_{t=1}^{M} \sum_{s=1}^{S} \sum_{s' \neq s}^{S} \| \mathbf{X_s} \mathbf{w_s^t} - \mathbf{X_{s'}} \mathbf{w_{s'}^t} \|^2, \\ T(\mathbf{w_s^t}) = \frac{1}{2} \sum_{t=1}^{M} \sum_{s=1}^{S} \| \mathbf{w_s^t} - \mathbf{w_s^{t+1}} \|^2 . \end{cases} \quad (5)$$

We set $\mathbf{w_s^{M+1}} = \mathbf{0}$ for the temporal progression term $T(\mathbf{w_s^t})$, since we only consider M time points in total. In addition, to facilitate the optimization of the temporal progression term, we restate $T(\mathbf{w_s^t})$ in an equivalent form as follows,

$$\sum_{t=1}^{M} \sum_{s=1}^{S} \| \mathbf{w_s^t} - \mathbf{w_s^{t+1}} \|^2$$
$$= \sum_{s=1}^{S} \| \mathbf{W_s H} \|^2 = \sum_{s=1}^{S} \| \sum_{t=1}^{M} \mathbf{w_s^t} \mathbf{h_t}^T \|^2, \quad (6)$$

where matrix $\mathbf{W_s} = [\mathbf{w_s^1}, \mathbf{w_s^2}, ..., \mathbf{w_s^M}] \in \mathbb{R}^{D_s \times M}$ and matrix $\mathbf{H} = [\mathbf{h_1}, \mathbf{h_2}, ..., \mathbf{h_T}]^T \in \mathbb{R}^{M \times (M-1)}$. The matrix $\mathbf{H}$ is precalculated by the following definition,

$$H_{ij} = \begin{cases} 1 & \text{if } i = j; \\ -1 & \text{if } i = j + 1; \\ 0 & \text{otherwise.} \end{cases} \quad (7)$$

The last two terms in Eqn.(3) penalize the generalization errors to avoid overfitting.

$$\begin{cases} \Omega(\mathbf{w_s^t}) = \frac{1}{2}\sum_{t=1}^{M}\sum_{s=1}^{S} \parallel \mathbf{w_s^t} \parallel^2, \\ \Psi(\boldsymbol{\alpha}) = \frac{1}{2} \parallel \boldsymbol{\alpha} \parallel^2 . \end{cases} \quad (8)$$

Overall, our proposed model has four parameters as shown in Eqn.(3). Parameters θ_1 and θ_2 respectively regularize the disagreement of different modalities for the same task and changes of health status between chronologically adjacent time points on the same modality. θ_3 and θ_4 regulate the strength of the l_2-norm regularization on multimodal multi-task learning and modality weights, respectively. We will detail the parameter tuning procedure in the experiments.

4.2 Alternative Optimization

We alternatively optimize $\boldsymbol{\alpha}$ and $\mathbf{w_s^t}$ via fixing the other one first. We will show that both of them admit analytic solutions.

4.2.1 Optimizing $\boldsymbol{\alpha}$ with $\mathbf{w_s^t}$ Fixed

We first fix $\mathbf{w_s^t}$ and optimize $\boldsymbol{\alpha}$. In such context, our objective function $\Phi(\mathbf{w_s^t}, \boldsymbol{\alpha})$ can be re-stated as $\Phi(\boldsymbol{\alpha})$,

$$\frac{1}{2}\sum_{t=1}^{M} \parallel \mathbf{y^t}\boldsymbol{e}^T\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\Lambda}\boldsymbol{\alpha} \parallel^2 + \frac{\theta_4}{2} \parallel \boldsymbol{\alpha} \parallel^2 + \xi(1 - \boldsymbol{e}^T\boldsymbol{\alpha}), \quad (9)$$

where $\boldsymbol{\Lambda} \in R^{D\times S}$ is a block matrix with its s-th diagonal block as $\mathbf{w_s^t}$; $\boldsymbol{e}$ is a vector with all elements being 1, and hence $\boldsymbol{e}^T\boldsymbol{\alpha}$ equals to 1; and ξ is the nonnegative Lagrange multiplier, which is frequently utilized in the optimization problem [33]; By setting the derivative of $\Phi(\boldsymbol{\alpha})$ with respect to $\boldsymbol{\alpha}$ to zero, we obtain,

$$\boldsymbol{\alpha} = \xi\boldsymbol{\Delta}^{-1}\boldsymbol{e}, \quad (10)$$

where $\boldsymbol{\Delta}$ is a positive definite and invertible matrix defined as,

$$\boldsymbol{\Delta} = (\mathbf{y}\boldsymbol{e}^T - \mathbf{X}\boldsymbol{\Lambda})^T(\mathbf{y}\boldsymbol{e}^T - \mathbf{X}\boldsymbol{\Lambda}) + \theta_4\mathbf{I} \quad (11)$$

In addition, based upon $\boldsymbol{e}^T\boldsymbol{\alpha} = 1$, we multiply both sides of Eqn.(10) with $\boldsymbol{e}^T$ to obtain,

$$\begin{cases} \xi = \frac{1}{\boldsymbol{e}^T\boldsymbol{\Delta}^{-1}\boldsymbol{e}}, \\ \boldsymbol{\alpha} = \frac{\boldsymbol{\Delta}^{-1}\boldsymbol{e}}{\boldsymbol{e}^T\boldsymbol{\Delta}^{-1}\boldsymbol{e}}. \end{cases} \quad (12)$$

4.2.2 Optimizing $\mathbf{w_s^t}$ with $\boldsymbol{\alpha}$ Fixed

We then fix $\boldsymbol{\alpha}$ and optimize $\mathbf{w_s^t}$. We take the derivative of $\Phi(\mathbf{w_s^t})$ with respect to $\mathbf{w_s^t}$,

$$\frac{\partial\Phi(\mathbf{w_s^t})}{\partial\mathbf{w_s^t}} = \alpha_s\mathbf{X_s}^T(\sum_{s=1}^{S}\alpha_s\mathbf{X_s}\mathbf{w_s^t} - \mathbf{y^t}) + \theta_1\mathbf{X_s}^T\sum_{s'\neq s}^{S}(\mathbf{X_s}\mathbf{w_s^t} - \mathbf{X_{s'}}\mathbf{w_{s'}^t}) + \theta_2\sum_{j=1}^{M}\mathbf{w_s^j}\mathbf{h_j}^T\mathbf{h_t} + \theta_3\mathbf{w_s^t}. \quad (13)$$

We set Eqn.(13) to zero and rearrange its terms. We arrive at the following equation,

$$\alpha_s\mathbf{X_s}^T\mathbf{y^t} = \left\{\alpha_s^2\mathbf{X_s}^T\mathbf{X_s} + \theta_1(S-1)\mathbf{X_s}^T\mathbf{X_s} + \theta_3\mathbf{I} + \theta_2\mathbf{h_t}^T\mathbf{h_t}\mathbf{I}\right\}\mathbf{w_s^t} + \sum_{s'\neq s}^{S}(\alpha_s\alpha_{s'} - \theta_1)\mathbf{X_s^T}\mathbf{X_{s'}}\mathbf{w_{s'}^t} + \theta_2\sum_{t'\neq t}^{M}\mathbf{h_t}^T\mathbf{h_{t'}}\mathbf{w_s^{t'}}, \quad (14)$$

where $\mathbf{I} \in \mathbb{R}^{D_s\times D_s}$ is an identity matrix. To ease the analysis, we align Eqn.(14) with the following form,

$$\mathbf{A_s^t} = \mathbf{B_s^t}\mathbf{w_s^t} + \sum_{s'\neq s}^{S}\mathbf{C_{ss'}}\mathbf{w_{s'}^t} + \sum_{t'\neq t}^{M}\mathbf{D^{tt'}}\mathbf{w_s^{t'}}. \quad (15)$$

By aligning Eqn.(14) with Eqn.(15), we derive the following set of equations,

$$\begin{cases} \mathbf{A_s^t} = \alpha_s\mathbf{X_s}^T\mathbf{y^t}; \\ \mathbf{B_s^t} = \alpha_s^2\mathbf{X_s}^T\mathbf{X_s} + \theta_1(S-1)\mathbf{X_s}^T\mathbf{X_s} + \theta_3\mathbf{I} + \theta_2\mathbf{h_t}^T\mathbf{h_t}\mathbf{I}; \\ \mathbf{C_{ss'}} = (\alpha_s\alpha_{s'} - \theta_1)\mathbf{X_s^T}\mathbf{X_{s'}}; \\ \mathbf{D^{tt'}} = \theta_2\mathbf{h_t}^T\mathbf{h_{t'}}\mathbf{I}. \end{cases} \quad (16)$$

Eqn.(15) and Eqn.(16) explicitly indicate that we must jointly learn $\mathbf{w_s^t}$ and $\mathbf{w_{s'}^{t'}}$ from a large set of equations, where $s' \neq s$ and $t' \neq t$. After combining the equations for all tasks on all modalities, we obtain a linear system as illustrated in Eqn.(17). It is worth noting that we can equivalently simplify this linear system as,

$$\mathbf{L}\mathbf{w} = \mathbf{a}, \quad (18)$$

where each entry in $\mathbf{L} \in \mathbb{R}^{(S\times M)\times(S\times M)}$ is a block matrix. Each block corresponds to a specific task on a specific modality, and its size is the dimensionality of the feature extracted from the corresponding modality. Similarly, $\mathbf{w}$ and $\mathbf{a}$ are block vectors with $S\times M$ blocks. So far, we have successfully transferred our proposed $\mathbf{aM^2L}$ model to a linear model. As we will show in the next subsection, $\mathbf{L}$ is positive definite, and thus invertible. Therefore, $\mathbf{w}$ can be obtained by solving the linear system shown in Eqn.(18).

4.2.3 Linear Model Demonstration

Before theoretically proving the invertible property of $\mathbf{L}$, we introduce three theorems.

Theorem 1: We consider multiple modalities and utilize S to denote the number of modalities. It is thus reasonable to assume that $S \geq 2$, and hence $\theta_1(S-1) \geq \theta_1$ when $\theta_1 > 0$.

Theorem 2: Without loss of generality, we assume $\mathbf{v_i}$ is an arbitrary block vector. We can then derive the following result,

$$\sum_{i=1}^{K}\mathbf{v_i}^T\mathbf{v_i} + \sum_{i=1}^{K}\sum_{j\neq i}^{K}\mathbf{v_i}^T\mathbf{v_j} = \frac{1}{2}\sum_{i=1}^{K} \parallel \mathbf{v_i} \parallel^2 + \frac{1}{2} \parallel \sum_{i=1}^{K}\mathbf{v_i} \parallel^2 \geq 0. \quad (19)$$

Theorem 3: Let's denote $\mathbf{z_i}$ as an arbitrary block vector. We will have the following inequality,

$$\sum_{i=1}^{K}\mathbf{z_i}^T\mathbf{z_i} - \sum_{i=1}^{K}\sum_{j\neq i}^{K}\mathbf{z_i}^T\mathbf{z_j} = \frac{1}{2} \parallel \mathbf{z_1} - \mathbf{z_K} \parallel^2 + \frac{1}{2}\sum_{i=2}^{K} \parallel \mathbf{z_i} - \mathbf{z_{i\text{-}1}} \parallel^2 \geq 0. \quad (20)$$

Next, we show that $\mathbf{L}$ is invertible by proving that $\mathbf{L}$ is a positive definite matrix. We define a $S \times M$ non-zero block vector, $\mathbf{g}^T = [\mathbf{g_{11}}^T, \mathbf{g_{21}}^T, ..., \mathbf{g_{S1}}^T, \mathbf{g_{12}}^T, ..., \mathbf{g_{st}}^T, ..., \mathbf{g_{1M}}^T, ..., \mathbf{g_{SM}}^T]$.

$$\begin{pmatrix} \mathbf{B_1^1} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} & \mathbf{D^{12}} & 0 & 0 & .. & 0 & \mathbf{D^{13}} & 0 & 0 & .. & 0 & .... & \mathbf{D^{1M}} & 0 & 0 & .. & 0 \\ \mathbf{C_{21}} & \mathbf{B_2^1} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} & 0 & \mathbf{D^{12}} & 0 & .. & 0 & 0 & \mathbf{D^{13}} & 0 & .. & 0 & .... & 0 & \mathbf{D^{1M}} & 0 & .. & 0 \\ \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & & \cdot & \cdot & \cdot & .. & \cdot \\ \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^1} & 0 & 0 & 0 & .. & \mathbf{D^{12}} & 0 & 0 & 0 & .. & \mathbf{D^{13}} & ... & 0 & 0 & 0 & .. & \mathbf{D^{1M}} \\ \mathbf{D^{21}} & 0 & 0 & .. & 0 & \mathbf{B_1^2} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} & \mathbf{D^{23}} & 0 & 0 & .. & 0 & .... & \mathbf{D^{2M}} & 0 & 0 & .. & 0 \\ 0 & \mathbf{D^{21}} & 0 & .. & 0 & \mathbf{C_{21}} & \mathbf{B_2^2} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} & 0 & \mathbf{D^{23}} & 0 & .. & 0 & ... & 0 & \mathbf{D^{2M}} & 0 & .. & 0 \\ \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & & \cdot & \cdot & \cdot & .. & \cdot \\ 0 & 0 & .. & 0 & \mathbf{D^{21}} & \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^2} & 0 & 0 & 0 & .. & \mathbf{D^{32}} & ... & 0 & 0 & .. & 0 & \mathbf{D^{2M}} \\ \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & & \cdot & \cdot & \cdot & .. & \cdot \\ \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & & \cdot & \cdot & \cdot & .. & \cdot \\ \mathbf{D^{M1}} & 0 & 0 & .. & 0 & \mathbf{D^{M2}} & 0 & 0 & .. & 0 & ... \; \mathbf{D^{M(M-1)}} & 0 & 0 & .. & 0 & \mathbf{B_1^M} & \mathbf{C_{12}} & \mathbf{C_{13}} & .. & \mathbf{C_{1S}} \\ 0 & \mathbf{D^{M1}} & 0 & .. & 0 & 0 & \mathbf{D^{M2}} & 0 & .. & 0 & ... \; 0 & \mathbf{D^{M(M-1)}} & 0 & .. & 0 & \mathbf{C_{21}} & \mathbf{B_2^M} & \mathbf{C_{23}} & .. & \mathbf{C_{2S}} \\ \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot & \cdot & \cdot & \cdot & .. & \cdot \\ 0 & 0 & .. & 0 & \mathbf{D^{M1}} & 0 & 0 & 0 & .. & \mathbf{D^{M2}} & ... \; 0 & 0 & 0 & .. & \mathbf{D^{M(M-1)}} & \mathbf{C_{S1}} & \mathbf{C_{S2}} & \mathbf{C_{S3}} & .. & \mathbf{B_S^M} \end{pmatrix} \begin{pmatrix} \mathbf{w_1^1} \\ \mathbf{w_2^1} \\ \cdot \\ \mathbf{w_S^1} \\ \mathbf{w_1^2} \\ \mathbf{w_2^2} \\ \cdot \\ \mathbf{w_S^2} \\ \cdot \\ \cdot \\ \mathbf{w_1^M} \\ \mathbf{w_2^M} \\ \cdot \\ \mathbf{w_S^M} \end{pmatrix} = \begin{pmatrix} \mathbf{A_1^1} \\ \mathbf{A_2^1} \\ \cdot \\ \mathbf{A_S^1} \\ \mathbf{A_1^2} \\ \mathbf{A_2^2} \\ \cdot \\ \mathbf{A_S^2} \\ \cdot \\ \cdot \\ \mathbf{A_1^M} \\ \mathbf{A_2^M} \\ \cdot \\ \mathbf{A_S^M} \end{pmatrix}. \quad (17)$$

We then derive that $\mathbf{g}^T\mathbf{L}\mathbf{g}$ equals to

$$= \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \mathbf{B_s^t} \mathbf{g_{st}} + \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \sum_{s' \neq s}^{S} \mathbf{C_{ss'}} \mathbf{g_{s't}} + \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \sum_{t' \neq t}^{M} \mathbf{D^{tt'}} \mathbf{g_{st'}}. \quad (21)$$

By substituting Eqn.(16) into Eqn.(21), we obtain $\mathbf{g}^T\mathbf{L}\mathbf{g}$,

$$= \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \Big\{ \alpha_s^2 \mathbf{X_s}^T \mathbf{X_s} + \theta_1 (S-1) \mathbf{X_s}^T \mathbf{X_s} + \quad (22)$$

$$\theta_3 \mathbf{I} + \theta_2 \mathbf{h_t}^T \mathbf{h_t} \mathbf{I} \Big\} \mathbf{g_{st}} + \sum_{s=1}^{S}\sum_{t=1}^{M} \sum_{s' \neq s}^{S} \mathbf{g_{st}}^T \Big\{ (\alpha_s \alpha_{s'} -$$

$$\theta_1) \mathbf{X_s^T X_{s'}} \Big\} \mathbf{g_{s't}} + \theta_2 \sum_{s=1}^{S}\sum_{t=1}^{M} \sum_{t' \neq t}^{M} \mathbf{g_{st}}^T \mathbf{h_t}^T \mathbf{h_{t'}} \mathbf{g_{st'}}. \quad (23)$$

As stated in ***theorem 1***, $\theta_1(S-1) \geq \theta_1$, we can thus further derive that $\mathbf{g}^T\mathbf{L}\ \mathbf{g}$ is greater than or equal to the followings,

$$\sum_{t=1}^{M} \Big\{ \sum_{s=1}^{S} \mathbf{g_{st}}^T (\alpha_s^2 \mathbf{X_s}^T \mathbf{X_s} + \theta_1 \mathbf{X_s}^T \mathbf{X_s}) \mathbf{g_{st}} + \sum_{s=1}^{S}\sum_{s' \neq s}^{S} \mathbf{g_{st}}^T$$

$$(\alpha_s \alpha_{s'} \mathbf{X_s}^T \mathbf{X_{s'}} - \theta_1 \mathbf{X_s}^T \mathbf{X_{s'}}) \mathbf{g_{s't}} \Big\} + \theta_2 \sum_{s=1}^{S} \Big\{ \sum_{t=1}^{M} \mathbf{g_{st}}^T \mathbf{h_t}^T \mathbf{h_t} \mathbf{g_{st}}$$

$$+ \sum_{t=1}^{M}\sum_{t' \neq t}^{M} \mathbf{g_{st}}^T \mathbf{h_t}^T \mathbf{h_{t'}} \mathbf{g_{st'}} \Big\} + \theta_3 \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \mathbf{g_{st}}. \quad (24)$$

Let's respectively denote block vector $\mathbf{v_s} = \alpha_s \mathbf{X_s} \mathbf{g_{st}}$, block vector $\mathbf{u_t} = \mathbf{h_t} \mathbf{g_{st}}$, and block vector $\mathbf{z_s} = \sqrt{\theta_1} \mathbf{X_s} \mathbf{g_{st}}$. We can rewrite the above formulas as follows,

$$= \sum_{t=1}^{M} \Big\{ \{ \sum_{s=1}^{S} \mathbf{v_s}^T \mathbf{v_s} + \sum_{s=1}^{S}\sum_{s' \neq s}^{S} \mathbf{v_s}^T \mathbf{v_{s'}} \} + \{ \sum_{s=1}^{S} \mathbf{z_s}^T \mathbf{z_s} -$$

$$\sum_{s=1}^{S}\sum_{s' \neq s}^{S} \mathbf{z_s}^T \mathbf{z_{s'}} \} \Big\} + \theta_2 \sum_{s=1}^{S} \Big\{ \sum_{t=1}^{M} \mathbf{u_t}^T \mathbf{u_t} + \sum_{t=1}^{M}\sum_{t' \neq t}^{M} \mathbf{u_t}^T \mathbf{u_{t'}} \Big\}$$

$$+ \theta_3 \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \mathbf{g_{st}}. \quad (25)$$

According to ***theorems 2*** and ***3***, we derive that Eqn.(25) is larger than or equal to,

$$\theta_3 \sum_{s=1}^{S}\sum_{t=1}^{M} \mathbf{g_{st}}^T \mathbf{g_{st}} = \theta_3 \sum_{s=1}^{S}\sum_{t=1}^{M} \parallel \mathbf{g_{st}} \parallel^2 > 0. \quad (26)$$

So far we have proven that $\mathbf{L}$ is a positive definite matrix. From the definition of positive definite matrix, we know that $\mathbf{L}$ is invertible.

4.3 Discussion

In the whole process of our alternative optimization, each step decreases the objective function value Φ, whose lower bound is zero and hence the convergence of our model is guaranteed [15].

The computational complexity of the training process is $O(T \times (O_1 + O_2))$, where O_1 and O_2 respectively equals to $(NS^2 + NDS + S^3)$ and (D^3M^3). T is the iteration times of the alternative optimization. S, N, M and D respectively refer to the number of modalities, patients, tasks, and the total feature dimensions over all the modalities. Usually, we consider less than 10 time points and modalities, hence both M and S are very small. D is in the order of a few hundreds. In our experiments, we studied less than one thousand patients in the training set. The process can be completed in less than five seconds excluding feature extraction on a desktop (3.4GHz and 16G memory). Therefore, our model can be easily scaled to other Web-scale or real-time applications, such as predicting the consumer taste progressions and event evolution modelling.

5. EXPERIMENTS

In this section, we verified our proposed model from various angles.

5.1 Experimental Settings

To facilitate the comparison among regression models, we employed the R-value to estimate the correlation coefficient between the predicted values and the ground truths [18],

$$R - value = \frac{\sum_{i=1}(p_i - \bar{p})(r_i - \bar{r})}{\sqrt{\sum_{i=1}(p_i - \bar{p})^2} \sqrt{\sum_{i=1}(r_i - \bar{r})^2}}, \quad (27)$$

where p_i is the predicted value and $\bar{p}$ is the average predicted value; r_i and $\bar{r}$ is the target value and the average target value, respectively. R-value always takes a value between -1 and 1, with 1 and -1 respectively indicating perfect positive and negative correlation. A correlation value close to 0 indicates no association between the variables. Besides, we also measured the regression performance by normalized

Mean Squared Error (nMSE) [29], which is the mean squared error divided by the variance of the target,

$$nMSE = \frac{\sum_{i=1}(p_i - r_i)^2}{\sum_{i=1}(r_i - \bar{r})^2}. \quad (28)$$

A better regression model usually has a higher R-value and a lower nMSE value.

Considering the small data size, the experimental results reported in this paper are based on 10-fold cross-validation. Note that patients with missing labels in the testing set are not utilized to assess our model, even their labels are estimated by our data completion method. Takeing "M36" as an example, 10-fold cross-validation assigns the testing set with approximately 81 patients. Thereinto, about 45 patients with real labels are taken for testing.

5.2 On Data Missing Issues

Due to the nature of longitudinal and periodic data collection, real chronic disease datasets usually contain a certain percentage of missing data, which greatly hampers the learning performance. Generally speaking, there exist two kinds of missing data. One is missing label (disease status), which is estimated by clinical measures, such as MMSE and ADAS-Cog. This is because some patients may be dead or absent at some time points. The other one is missing modality. In particular, the patients may not provide their complete health modalities at the baseline owing to privacy and security concerns. In fact, Tables 2 and 3 respectively show the details of missing modality and missing label in our dataset.

On the other hand, the mature matrix factorization technique that is able to deal with missing data in a principled way, has attracted a lot of attentions from diverse communities [4, 13, 3]. Basically, matrix factorization factorizes a given matrix into two latent matrices, such that their multiplication approximates the original one. The entries in the two latent matrices are inferred by the observed values in the original matrix only, and overfitting is avoided through an appropriate regularized model. In this work, we merge modality matrix and the disease status matrix as the original matrix $\mathbf{O} = [\mathbf{X}, \mathbf{Y}] \in \mathbb{R}^{N \times (D+M)}$, and aim to infer the missing entries in $\mathbf{O}$. Based on this framework[10], we implemented a flexible learning rate adjuster to monitor and adjust the learning rate in matrix factorization. This adjuster is triggered on each epoch. It will shrink the learning rate if the value of the objective function goes up. The idea is that in this case the learning algorithm is overshooting the bottom of the objective function. On the other hand, the adjuster will increase the learning rate if the value of the objective function decreases too slowly. This makes our learning rate parameter less dependent to the initial value. Though it is not a very mathematically principled approach, it works well in practice and really accelerates the learning process.

To examine the necessity, effectiveness and efficiency of our proposed data completion method, we evaluated our **aM^{2}L** model under the following settings.

- **REM**: We simply removed the patients with either missing modality or missing label.
- **NN**: We replaced the missing values in $\mathbf{O}$ with the corresponding values from the nearest column based on META and MRI modality. Gaussian kernel function [20] was employed to estimate the pairwise similarity.
- **MF**: The missing data was completed by a matrix factorization method. The learning rate was set as a sufficient small value of 0.00001 to avoid oscillation. This value was fixed across the training process.
- **aMF**: The missing data was inferred by a matrix factorization method with adaptive learning rate adjuster. The initial learning rate was set as 0.01.

[10] http://www.quuxlabs.com/blog/2010/09/matrix-factorization-a-simple-tutorial-and-implementation-in-python/

Table 5: Performance comparison among different data completion methods. Time refers to the time cost of data completion in terms of epoch.

Method	Predicted by MSMT Learning				Time
	MMSE		ADAS-Cog		
	R-value	nMSE	R-value	nMSE	
DEL	0.6587	0.1736	0.6944	0.0602	–
KNN	0.8597	0.1375	0.8784	0.0413	–
MF	0.8901	0.1181	0.9168	0.0318	3174
aMF	0.8901	0.1181	0.9168	0.0318	32

In Table 5, we show the performance of our model with different data completion methods. From this table, we have the following observations: 1) nMSE under ADAS-Cog is much smaller as compared against that under MMSE. That is because the variance of ADAS-Cog is very large. ADAS-Cog scores range from 0-70, while MMSE scores range from 0-30. But this does not affect the comparison among models under the same cognitive score. 2) **REM** achieves the worst performance across different cognitive measures and evaluation criteria. A possible reason is that it substantially reduces the training size and patients with incomplete data cannot be leveraged to train the regression model. According to our statistic, only 184 patients have all the five modalities and 429 patients have all the labels at M06, M12, M24 and M36. Worse still is only 93 patients having all the labels and five sources simultaneously. This indicates that the data completion procedure is critical. 3) **aMF** exhibits better performance than **NN** and **MF**, in terms of effectiveness and efficiency, respectively. This is because **NN** is unable to explore global information for missing data inference, and **MF** is unable to flexibly adjust its learning rate and thus it takes a long time to reach convergence.

5.3 On Parameter Tuning

We have four key parameters as shown in Eqn.(3). The optimal values of these parameters were carefully tuned with 5-fold cross-validation in the training data. In particular, based upon the 10-fold cross-validation, we have around 736 patients during each training round. We performed 5-fold cross-validation on the 736 patients to learn the optimal parameters by grid search between 10^{-2} to 10^2 with small but adaptive step size. In particular, the step size was 0.01, 0.05, 1, and 5 for the range of [0.01,0.1], [0.1,1], [1,10] and [10,100], respectively. The parameters corresponding to the best R-value were used to report the final results. For other competitors, the procedures to tune the parameters are analogous to ensure fair comparison.

5.4 On Component-wise Analysis

To verify how effective each component is in the proposed **aM^{2}L** model, we compared the performance of the following

Table 6: Effectiveness evaluation of each involved component in our proposed model.

Methods	MMSE		ADAS-Cog		P-value
	R-value	nMSE	R-value	nMSE	
Ridge	0.7982	0.1577	0.8485	0.0433	3.2e-12
noTP	0.8245	0.1496	0.8613	0.0429	4.8e-10
noMA	0.8706	0.1275	0.9017	0.0362	5.9e-4
M^{2}L	0.8794	0.1233	0.9094	0.0347	1.1e-3
aM^{2}L	**0.8901**	**0.1181**	**0.9168**	**0.0318**	–

methods, which can be derived from our model by removing some terms:

- **Ridge**: Ridge regression is a simple approach to estimate the future health statuses by modelling modalities and tasks at different time points separately [8]. It is typically formulated as $(\mathbf{y^t}-\mathbf{Xw^t})^2+\delta \parallel \mathbf{w^t} \parallel_F^2$, and admits an analytical solution given by $\mathbf{w^t} = (\mathbf{X}^T\mathbf{X} + \delta\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$. It is a special case of our model, when $\theta_1 = 0$, $\theta_2 = 0$, $\theta_4 = 0$, and $\alpha_s = 0.2$ ($\alpha_s = \frac{1}{S}$ and $S = 5$).
- **noTP**: We did not consider temporal progression by setting $\theta_2 = 0$.
- **noMA**: We did not take modality agreement into consideration by setting $\theta_1 = 0$.
- **M^{2}L**: Our proposed regression model without considering the confidence of each modality. This can be derived by setting $\theta_4 = 0$ and $\alpha_s = \frac{1}{S}$.
- **aM^{2}L**: Our proposed regression model.

The results of component-wise analysis are reported in Table 6. From this table, the following observations can be made: 1) **Ridge** achieves the most unsatisfactory results. This may be caused by the fact that **Ridge** neither explores the temporal smoothness among tasks, nor considers the consistency relatedness among modalities. This results imply that jointly modelling the dual-heterogeneities of disease progression is of vital importance: the prediction task at each time point has features from multiple modalities, and multiple tasks are related to each other in chronological order. 2) **noMA** and **M^{2}L** perform better than **noTP**. This demonstrates that multi-task learning has even more encouraging effects as compared to multimodal analysis. This is because multi-task learning effectively increases the number of samples by learning multiple related tasks simultaneously and our data size is not very large. 3) **aM^{2}L** is superior to others. This further demonstrates that every involved component in our proposed model is indispensable. For instance, as compared to **M^{2}L**, **aM^{2}L** shows its advantage of adaptive modality confidences.

We also conducted the analysis of variance (known as ANOVA) based on R-value under MMSE. In particular, we performed paired t-test between our **aM^{2}L** model and each of the benchmarks based on 10-fold cross validation. We found that all the p-values are much smaller than 0.05, which shows that the improvements of our proposed model are statistically significant.

5.5 On Model Performance Comparison

Beside component analysis, we carried out experiments to compare the overall effectiveness of our proposed **aM^{2}L** model with other state-of-the-arts regression models in the disease progression domain, and other multi-view multi-task learning models:

- **TGL**: A Temporal Group Lasso model proposed in [35] to predict the chronic disease progression. It captures the relatedness among tasks at sequential time points by a temporal group lasso regularizer.
- **cFSGL**: A novel Convex Fused Sparse Group Lasso formulation introduced in [34] to model the disease progression. It simultaneously selects task-shared and task-specific features using the sparse group Lasso penalty.
- **nFGL**: A non-Convex Fused Group Lasso method proposed to model the disease progression in Eqn.17 of [34]. It is a composite $l_{0.5,1}$-like penalty. The difference of convex programming technique was employed to solve the non-convex formulations.
- **regMVMT**: A semi-supervised inductive multi-view multi-task learning model introduced in [32]. As reported in [32], this model outperforms most of the prevailing multi-view multi-task learning approaches. That is why we did not selected other multi-view multi-task learning models as competitors.

Table 7 shows the performance of various disease progression models in terms of nMSE and R-value. From this table, we observed the following points: 1) Our proposed **aM^{2}L** model is remarkably and consistently better than the first three prevailing disease progression models. The reason could be that none of the disease progression models jointly leverage the prior knowledge of modality agreement and their confidences while modelling the task relatedness. This observation manifests the consistent relations among modalities and their confidences are able to appropriately reinforce the descriptions of health conditions and hence enhance the prediction performance of disease progression. 2) **aM^{2}L** outperforms **regMVMT**. This is because the latter does not flexibly consider the modality weights and is unable to make full advantage of the labeled samples. Meanwhile, it assumes that the relations among tasks are uniformly distributed, which is often not true in the real world. 3) Some learning models on some tasks, where the real labeled samples are relatively less, surprisingly achieve a better performance. For example, the learning performance on M24 is greater than that on M06. Though we estimated the missing labels before the training, there exists a certain bias between the estimated labels and the true labels. After inspecting the distribution of patients, we found it is reasonable. The percentage of NL changed from 28.3% in M06 to 31.4% in M24, which results in a smaller variance and hence a possible larger nMSE.

5.6 On Multimodal Analysis

We also studied the effectiveness of different modality combination. Recall that each patient after data completion is represented by five modalities in our work. Instead of exhaustively examining all possible modality combinations, we fed the following representative combinations into our proposed model and validated their description power:

- **IMG**: Imaging set comprises of MRI and PET.
- **nIMG**: Non-imaging set has CSF, PROT, and META.
- **ALL**: We incorporated all the five modality into our model simultaneously.

Table 7: Comparison of our proposed model with other state-of-the-art disease regression models and multi-view multi-task learning models on longitudinal MMSE and ADAS-Cog prediction.

Models	MMSE						ADAS-Cog					
	All Tasks		Individual Task (nMSE)				All Tasks		Individual Task (nMSE)			
	R-value	nMSE	M06	M12	M24	M36	R-value	nMSE	M06	M12	M24	M36
TGL	0.8464	0.1416	0.1610	0.1485	0.1072	0.1459	0.8793	0.0412	0.0408	0.0380	0.0412	0.0469
cFSGL	0.8523	0.1403	0.1586	0.1479	0.1073	0.1434	0.8906	0.0389	0.0406	0.0356	0.0377	0.0430
nFGL	0.8557	0.1382	0.1581	0.1462	0.1041	0.1394	0.8928	0.0383	0.0407	0.0352	0.0370	0.0409
regMVMT	0.8687	0.1284	0.1467	0.1324	0.0991	0.1315	0.8975	0.0363	0.0381	0.0344	0.0356	0.0399
aM^{2}L	**0.8901**	**0.1181**	**0.1313**	**0.1159**	**0.0804**	**0.1143**	**0.9168**	**0.0318**	**0.0332**	**0.0298**	**0.0309**	**0.0338**

Table 8: Performance comparison among different multimedia and multi-modal combinations.

Sources	Predicted by MSMT Learning				p-value
	MMSE		ADAS-Cog		
	R-value	nMSE	R-value	nMSE	
IMG	0.8694	0.1281	0.8993	0.0375	1.3e-5
nIMG	0.8804	0.1205	0.9081	0.0353	2.6e-3
ALL	**0.8901**	**0.1181**	**0.9168**	**0.0318**	–

Table 8 summarizes the multimodal analysis and the paired t-test results. We observed that: 1) The model trained on **nIMG** outperforms that trained on **IMG**. This tells us the non-imaging modality combination is much more informative as compared against the imaging scans. 2) The paired t-test results show that the model on all the five modalities is significantly better than those trained on others. This implies that the visual features from imaging set can enhance the learning performance. 3) We examined the modality confidence α_s during training on **ALL**. We found that the weight on META, PROT, MRI, PET, CSF in turn goes down. This reveals that META is the most discriminative modality for AD characterization.

6. CONCLUSION AND FUTURE WORK

In this paper, we presented a new approach to model the progression of chronic diseases based on multimedia and multimodal observational data. It seamlessly unifies the modality agreement, temporal progression and the modality confidences. Extensive experiments on a publicly acquirable dataset have well verified the effectiveness of our proposed model and each of its components.

In future, we plan to further improve the model in the following directions. First of all, we will further consider the structural relations between tasks, instead of just linear relations. Secondly, beyond modality agreement, we will extend our model to incorporate the complementary relations among modalities. At last, we plan to apply our model to other diseases.

7. ACKNOWLEDGEMENTS

This research is supported by the Singapore National Research Foundation under its International Research Centre @ Singapore Funding Initiative and administered by the IDM Programme Office.

8. REFERENCES

[1] A. Bogomolov, B. Lepri, M. Ferron, F. Pianesi, and A. S. Pentland. Daily stress recognition from mobile phone data, weather conditions and individual traits. In *ACM MM*, 2014.

[2] Y. F. C. Misra and C. Davatzikos. 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*, 2009.

[3] C.-M. Chen, M.-F. Tsai, J.-Y. Liu, and Y.-H. Yang. Using emotional context from article for contextual music recommendation. In *ACM MM*, 2013.

[4] P. Cui, Z. Wang, and Z. Su. What videos are similar with you?: Learning a common attributed representation for video recommendation. In *ACM MM*, 2014.

[5] R. Desikan, H. Cabral, F. Settecase, C. Hess, W. Dillon, C. Glastonbury, M. Weiner, N. Schmansky, D. Salat, and B. Fischl. Automated mri measures predict progression to alzheimer's disease. *Neurobiology of aging*, 2010.

[6] S. Duchesne, A. Caroli, C. Geroldi, D. Collins, and G. Frisoni. Relating one-year cognitive change in mild cognitive impairment to baseline mri features. *NeuroImage*, 2009.

[7] S. Ebadollahi, A. R. Coden, M. A. Tanenblatt, S.-F. Chang, T. Syeda-Mahmood, and A. Amir. Concept-based electronic health records: Opportunities and challenges. In *ACM MM*, 2006.

[8] M. Eliot, J. Ferguson, M. P. Reilly, and A. S. Foulkes. Ridge regression for longitudinal biomarker data. *JMLR*, 2011.

[9] J. He and R. Lawrence. A graph-based framework for multi-task multi-view learning. In *ICML*, 2011.

[10] K. Ito, B. Corrigan, Q. Zhao, J. French, R. Miller, H. Soares, E. Katz, T. Nicholas, B. Billing, R. Anziano, and T. Fullerton. Disease progression model for cognitive deterioration from alzheimer's disease neuroimaging initiative database. *Alzheimer's and Dementia*, 2010.

[11] X. Jin, F. Zhuang, S. Wang, Q. He, and Z. Shi. Shared structure learning for multiple tasks with multiple views. In *MLKDD*. 2013.

[12] P. JR, C. RE, and D. PM. Neuroimaging and early diagnosis of alzheimer disease: a look to the future. *Radiology*, 2003.

[13] Z. Li, J. Liu, X. Zhu, T. Liu, and H. Lu. Image annotation using multi-correlation probabilistic matrix factorization. In *ACM MM*, 2010.

[14] H. Lin, J. Jia, Q. Guo, Y. Xue, Q. Li, J. Huang, L. Cai, and L. Feng. User-level psychological stress detection from social media using deep neural network. In *ACM MM*, 2014.

[15] Z. Ma, Y. Yang, F. Nie, J. Uijlings, and N. Sebe. Exploiting the entire feature space with sparsity for automatic image annotation. In *ACM MM*, 2011.

[16] L. Nie, M. Akbari, T. Li, and T.-S. Chua. A joint local-global approach for medical terminology assignment. In *SIGIR, MIR Workshop*, 2014.

[17] L. Nie, T. Li, M. Akbari, J. Shen, and T.-S. Chua. Wenzher: Comprehensive vertical search for healthcare domain. In *SIGIR*, 2014.

[18] L. Nie, M. Wang, Z.-J. Zha, and T.-S. Chua. Oracle in image search: A content-based approach to performance prediction. *TOIS*, 2012.

[19] L. Nie, M. Wang, L. Zhang, S. Yan, B. Zhang, and T.-S. Chua. Disease inference from health-related questions via sparse deep learning. *TKDE*, 2015.

[20] L. Nie, S. Yan, M. Wang, R. Hong, and T.-S. Chua. Harvesting visual concepts for image search with complex queries. In *ACM MM*, 2012.

[21] L. Nie, Y.-L. Zhao, M. Akbari, J. Shen, and T.-S. Chua. Bridging the vocabulary gap between health seekers and healthcare knowledge. *TKDE*, 2015.

[22] R. K. Pearson, R. J. Kingan, and A. Hochberg. Disease progression modeling from historical clinical databases. In *SIGKDD*, 2005.

[23] C. Stonnington, C. Chu, S. Klöppel, C. Jack, J. Ashburner, R. Frackowiak, and the Alzheimer Disease Neuroimaging Initiative. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *NeuroImage*, 2010.

[24] N. Tzourio-Mazoyer, B. Landeau, D. Papathanassiou, F. Crivello, O. Etard, N. Delcroix, B. Mazoyer, and M. Joliot. Automated anatomical labeling of activations in spm using a macroscopic anatomical parcellation of the mni mri single-subject brain. *NeuroImage*, 2002.

[25] S. Xiang, L. Yuan, W. Fan, Y. Wang, P. M. Thompson, and J. Ye. Multi-source learning with block-wise missing data for alzheimer's disease prediction. In *SIGKDD*, 2013.

[26] Q. Xu, Q. Huang, and Y. Yao. Online crowdsourcing subjective image quality assessment. In *ACM MM*, 2012.

[27] X. Yang, S. Kim, and E. P. Xing. Heterogeneous multitask learning with joint sparsity constraints. In *NIPS*, 2009.

[28] J. Ye, K. Chen, T. Wu, J. Li, Z. Zhao, R. Patel, M. Bae, R. Janardan, H. Liu, G. Alexander, and E. Reiman. Heterogeneous data fusion for alzheimer's disease study. In *SIGKDD*, 2008.

[29] Z. Yu and D.-Y. Yeung. Multi-task learning using generalized t process. *JMLR*, 2010.

[30] L. Yuan, Y. Wang, P. M. Thompson, V. A. Narayan, and J. Ye. Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data. In *SIGKDD*, 2012.

[31] D. Zhang and D. Shen. Multi-modal multi-task learning for joint prediction of clinical scores in alzheimer's disease. In *Multimodal Brain Image Analysis*. 2011.

[32] J. Zhang and J. Huan. Inductive multi-task learning with multiple view data. In *SIGKDD*, 2012.

[33] X. Zhao, G. Li, M. Wang, J. Yuan, Z.-J. Zha, Z. Li, and T.-S. Chua. Integrating rich information for video recommendation with multi-task rank aggregation. In *ACM MM*, 2011.

[34] J. Zhou, J. Liu, V. A. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *SIGKDD*, 2012.

[35] J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *SIGKDD*, 2011.