Data-Driven Longitudinal Modeling and Prediction of Symptom Dynamics in Major Depressive Disorder: Integrating Factor Graphs and Learning Methods

Arjun P. Athreya*, Subho S. Banerjee*, Drew Neavin†, Rima Kaddurah-Daouk‡, A. John Rush‡, Mark A. Frye†, Liewei Wang†, Richard M. Weinshilboum†, William V. Bobo† and Ravishankar K. Iyer*
*University of Illinois at Urbana-Champaign, USA. Email: {athreya2, ssbaner2, rkiyer}@illinois.edu
†Mayo Clinic, USA. Email: {neavin.drew, mfrye, wang.liewei, weinshilboum.richard, bobo.william}@mayo.edu
‡Duke University, USA. Email: {john.rush, rima.kaddurahdaouk}@duke.edu

***Abstract*—This paper proposes a data-driven longitudinal model that brings together factor graphs and learning methods to demonstrate a significant improvement in predictability in clinical outcomes of patients with major depressive disorder treated with antidepressants. Using data from the Mayo Clinic PGRN-AMPS trial and the STAR*D trial for validation, this work makes two significant contributions in the context of predictability in psychiatric therapeutic outcomes. *First*, we establish symptom dynamics in response to antidepressants by using the forward algorithm on a factor graph. Symptom dynamics are the changes in the symptom severity that are most likely to occur because of the antidepressants taken during the trial, and the associated clinical outcomes at 4 weeks and 8 weeks into the trial. The structure of the factor graph is inferred by using unsupervised learning to stratify patients by the similarity of their overall symptom severity. *Second*, by using metabolomics data as an accurate biological measure in addition to symptom survey data and other patient history information, the prediction of clinical outcomes such as *response* and *remission* significantly improved from 30% to 68% in men, and from 35% to 72% in women. This work demonstrates a significant difference in how men and women respond to antidepressants in terms of their symptom dynamics, and also shows that top predictors of clinical outcomes for men and women are significantly different and known to play a role in behavioral sciences.**

1. Introduction

Major depressive disorder (MDD) affects over 350 million patients worldwide [1]–[3]. Antidepressant medications such as selective serotonin reuptake inhibitors (SSRIs) are the primary options for pharmacotherapy in adults with MDD [4]. It is known that baseline data (prior to therapy) that consist of 1) *social and demographic data* (S in Table 1) and 2) *clinical data* comprising the patient's responses to questionnaires as assessed by a clinician (C in Table 1) do not have sufficient predictive validity to guide clinical decision-making [4]–[9]. Unlike other major diseases, such as cancer or diabetes, MDD currently has no validated biomarkers or other indicators that can be used to predict antidepressant treatment outcomes. To further the understanding of the pathophysiology of MDD, plasma *metabolomic* concentrations from patients during three stages (at baseline, 4 weeks, and 8 weeks) of the trial was collected in the Mayo Clinic Pharmacogeomics Research Network Antidepressant Medical Pharmacogenomic Study (Mayo PGRN-AMPS) clinical trial [10] (the largest single-center trial in the USA), in addition to social, demographic and clinical data. Using these biological measures and existing validated clinical measures for 603 patients $\times\, 65$ variables collected across three time-points in the Mayo PGRN-AMPS trial, this work addresses two key questions: 1) what are the longitudinal dynamics of symptoms in response to antidepressants, i.e., what changes to the symptom severity are most likely to occur because of the antidepressants taken during the treatment, and the associated clinical outcomes at 4 weeks and 8 weeks into the trial? and 2) how much would the integration of biological measures with validated clinical measures improve the predictability in clinical outcomes of patients treated with antidepressants? Data from the STAR*D trial [11], which is the largest multi-center trial in the USA, are used for validation of findings from the Mayo PGRN-AMPS trial. While previous work has looked at integrating biological

TABLE 1: Data ($D = [S : C : B]$)

Total Patients: 603. Men: Total: 222. With metabolomics: 99. Women: Total: 381. With metabolomics: 191.
Social and demographic data (S) collected only at baseline
Age (in years)
Body mass index (BMI in kg/m^2)
Depression in {parents, siblings, children}
Bipolar disorder in {parents, siblings, children}
Alcohol abuse by {parents, siblings, children}
Drug abuse by {parents, siblings, children}
Seasonal pattern in symptom occurrence
History of psychotherapy
Clinical data (C)
Clinician-rated Quick Inventory of Depressive Symptomatology (QIDS-C) questionnaire (16 questions)
QIDS-C total score
Biological data (B)
31 metabolites from the HPLC LCECA platform

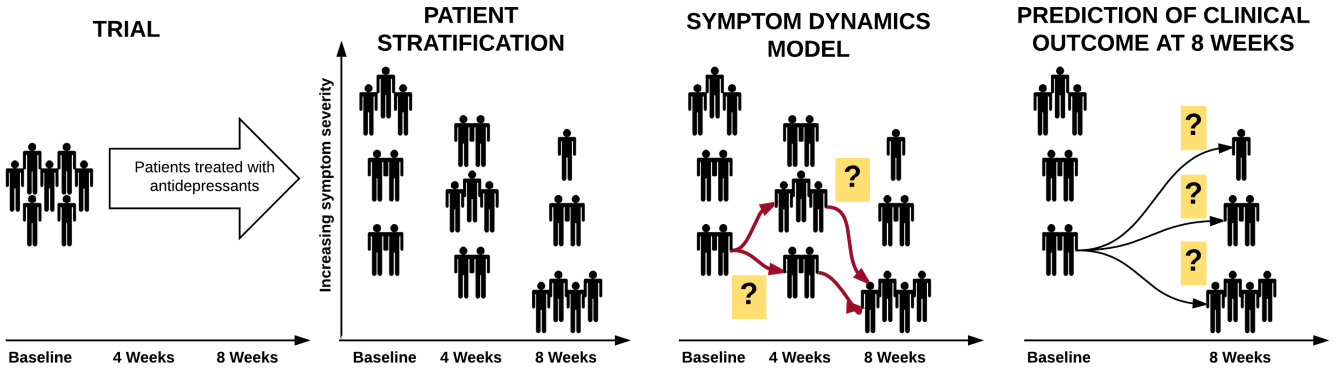


Figure 1: The proposed analyses to study symptom dynamics and predictability of clinical outcomes at 8 weeks.

data of multiple modalities to predict outcomes [12]–[15], there remains an open problem of how biological measures can be integrated with a patient's response to a questionnaire in the absence of existing biomarker knowledge.

The following are the main contributions of this work as illustrated in Fig 1.

1) **Patient stratification** (Sec. 3): Model-based unsupervised learning identified three distinct clusters of men and women at all three time-points in the Mayo PGRN-AMPS trial based on their overall symptom severity. The patient stratification was statistically similar in the STAR*D trial. The relevance of the clustering behavior is seen at 8 weeks, at which point the three clusters comprised all patients who 1) achieved *remission*, 2) demonstrated *response* but not *remission*, and 3) demonstrated no response, and no remission respectively, which agrees with the definitions of clinical outcomes in psychiatry.
2) **Longitudinal symptom dynamics model** (Sec. 4): A factor graph [16], [17] was developed to capture the relationships between clusters of patients at consecutive time-points of the trial and associated variables, such as clinical outcomes, metabolomics and demographic data. The choice of factor graphs was driven by their ability to provide a compact expressive representation of random variables and to subsume Bayesian networks, Markov random fields (MRFs), and hidden Markov models [16], [17]; further, they have been shown to be effective in modeling longitudinal electronic health records data of diabetic patients [18]. For a given level of symptom severity prior to the treatment, the *forward algorithm* on the inferred factor graph helped establish the symptom dynamics. The symptom response to antidepressants was observed to be identical in both trials, and the key finding was that symptom dynamics were significantly different in men and women.
3) **Improved prediction of clinical outcomes by integrating metabolomics data** (Sec. 5): This work demonstrates that for both *response* and *remission*, the addition of baseline metabolomics data to social, demographic, and clinical data significantly increased the predictability of clinical outcomes from 30% to 68% in men, and from 35% to 72% in women. Metabolites whose baseline concentrations were correlated with the symptom severity at 8 weeks were chosen, then normalized along with clinical and demographic data for training supervised learning methods. The key finding is that the top-predictors from the prediction model are significantly different in men and women, and these predictors are known to be associated with clinical outcomes in psychiatry.

2. Data

The Mayo PGRN-AMPS trial was designed to assess the clinical outcomes of adults (aged 18–84 years) with non-psychotic MDD after 4 and 8 weeks of open-label treatment with citalopram or escitalopram and to examine metabolomic and genomic factors associated with those outcomes [10]. Subjects were recruited from primary and specialty care settings in and near Rochester, MN from March 2005 to May 2013. All psychiatric diagnoses were confirmed at the screening visit using modules of the Structured Clinical Interview for DSM-IV (SCID) administered by trained clinical research staff. The data $D = [S : C : B]$ analyzed in this work comprise social and demographic variables (S), clinical measures (C), and biological measures (B) are tabulated in Table 1. The social and demographic data (S) were assessed only at baseline. The treatment outcomes were established using the 16-item, clinician-rated version of the Quick Inventory of Depressive Symptomatology (QIDS-C) at baseline, 4 weeks and 8 weeks; the results comprised the clinical data C, which included the responses to the 16 QIDS-C questions and the total QIDS-C score of the symptom severity [19]. The biological data B comprised 31 metabolites from samples collected at baseline, 4 weeks, and 8 weeks. Samples were assayed on a high-performance liquid chromatography (HPLC) electrochemical coulometric array (LCECA) platform to obtain the standardized measures of concentrations of metabolites.

Clinical definitions: *Response* is defined as a 50% reduction in baseline symptoms as measured at 4 weeks or 8 weeks. If the total QIDS-C score measured at 8 weeks is ≤ 5, then the patient is said to have achieved *remission*.

3. Patient Stratification

To study the longitudinal behavior of depressive symptoms in response to antidepressants given a patient's severity at baseline, an understanding of how symptoms broadly change in the trial is needed. To gain this understanding, patients were stratified (clustered) based on their total QIDS-C score, which is a measure of the severity of depression symptoms in patients. The clusters will then help guide the analytics needed to study the likely symptom dynamics during the trial in a patient given his or her starting symptom severity. Based on existing knowledge of gender differences in response to antidepressants [20], the patients were stratified separately by gender. Currently, there are no established mechanisms in which patients with MDD are stratified and has been limited by lack of access to data from large trials.

Observation: The p-value from the Shapiro-Wilk test of the the total QIDS-C score from all three time-points of the trial and in both men and women, was less than the significance level ($\alpha = 0.05$). This meant that the symptom severity scores were not normally distributed, as we rejected the null hypothesis of the Shapiro-Wilk test (i.e., that the data are normally distributed).

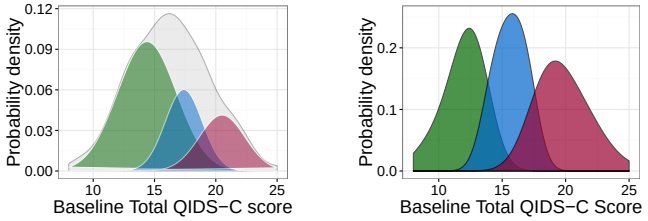
Approach: The fact that symptom severity are not normality distributed meant that the k-means clustering algorithm would not suitable as a clustering algorithm here. Without a loss in generality, under the assumption that the data (x: total QIDS-C score) was distributed as a mixture of Gaussians (referred to as a Gaussian mixture model (GMM)), we developed the patient stratification workflow (Algorithm 1). Starting with an assumption that the data have at least two components in the GMM, we used the expectation maximization (EM) algorithm to estimate the sufficient statistics parameters of the Gaussian components (mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\sigma}^2$) of the GMM as shown in Fig. 2(a). $10,000$ samples were randomly drawn from the inferred distributions (generateSamples). Next, the Kolmogorov-Smirnov test (ks.test) was used to test whether the distribution of the generated data was statistically similar to that of the original data. If the p-value (p) was less than the significance level ($\alpha = 0.05$), then we would reject the null-hypothesis that the two distributions are not similar. If that happened, the number of components was increased by one, and tested for similarity in the two distributions. Once the minimum number of components K in the GMM was obtained for which the generated and input data's distributions are similar, K clusters $\mathcal{C} = \{C^k; \forall k \in 1:K\}$ ordered by the increasing mean (μ_k) of the components were the outputs of the workflow [21]. Patients were assigned to the component that maximizes the likelihood $\mathcal{L}(x)$ given the component's sufficient statistics (gmmCluster), illustrated in Fig. 2(b) and described by Equation 1

$$\underset{k\in[1:K]}{\operatorname{argmax}}\, \mathcal{L}_k(x) \text{ where } \mathcal{L}_k(x) = \mathcal{N}(x, \mu_k, \sigma_k^2). \quad (1)$$

Results: At each time-point $t \in \{b(\text{baseline}), f(4 \text{ weeks}), e(8 \text{ weeks}) \}$, we found three clusters of men and women as shown in Fig. 3(a), in which the pie chart for each cluster was positioned at the mean of the cluster's average symptom severity score. Clusters at the baseline are $\mathcal{C}_b = \{\mathcal{C}_b^1, \mathcal{C}_b^2, \mathcal{C}_b^2\}$, at 4 weeks are $\mathcal{C}_f = \{\mathcal{C}_f^1, \mathcal{C}_f^2, \mathcal{C}_f^2\}$, and at 8 weeks are $\mathcal{C}_e = \{\mathcal{C}_e^1, \mathcal{C}_e^2, \mathcal{C}_e^2\}$. The composition of the clusters at 4 weeks and 8 weeks based on the baseline cluster assignments is captured in Fig. 3(a). It can be seen that a significant majority of patients (96% of the 603) show reduction in their symptoms at 4 weeks and 8 weeks. The clinical value of the clustering behavior is that $\mathcal{C}_e^1$ in both men and women captures all patients who achieved *remission* at the end of 8 weeks. Furthermore, the $\mathcal{C}_e^2$ in both men and women comprised patients who demonstrated *response* but did not achieve *remission*. Finally, patients in $\mathcal{C}_f^3$ in both men and women did not exhibit *response* or achieve *remission*. The same workflow demonstrated identical patient stratification in the STAR*D trial, i.e., the Kolmogorov-Smirnov test for symptom severity scores between clusters of similar average symptom severity had p-value > 0.8. From the analytics perspective, *the significance of the replication of patient stratification in two of the largest trials is that the clustering behavior followed the existing definitions of clinical outcomes in psychiatry.* Patient stratification as we will see next, lays the foundation to model the symptom dynamics.

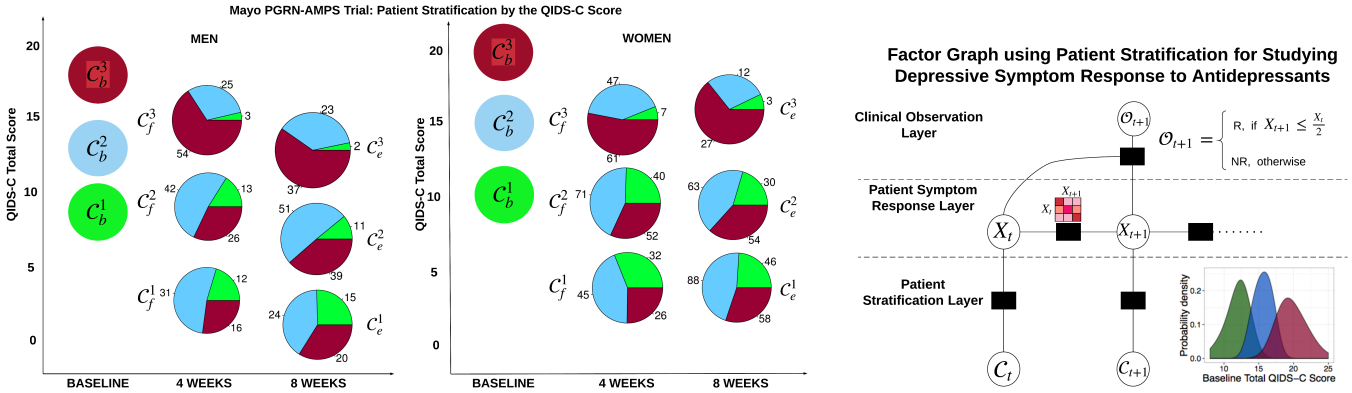
Algorithm 1 Patient stratification

Input: $x \leftarrow$ Total QIDS-C Scores
1: $k \leftarrow 2$
2: $\mathcal{C} \leftarrow \emptyset$
3: $\alpha \leftarrow 0.05$
4: $p \leftarrow 0$
5: **while** $p \leq \alpha$ **do**
6: $\quad \{\boldsymbol{\mu}, \boldsymbol{\sigma^2}\} \leftarrow$ EM(x, k)
7: $\quad x' \leftarrow$ generateSamples $(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
8: $\quad p \leftarrow$ ks.test(x, x')
9: $\quad$ **if** $p >$ significanceLevel **then**
10: $\quad\quad \mathcal{C} \leftarrow$ gmmCluster$(\boldsymbol{\mu}, \boldsymbol{\sigma^2})$
11: $\quad$ **end if**
12: $\quad k \leftarrow k + 1$
13: **end while**
Output: $\mathcal{C}$



(a) Estimating Distributions (b) Clusters from inferred GMM

Figure 2: Fig. (a) illustrates the inference of mixtures comprising the distribution of symptom severity scores. Fig. (b) illustrates the clusters inferred using the sufficient statistics of components inferred in Fig. (a).



(a) Patient stratification (b) Factor graph

Figure 3: Fig. (a) is the illustration of the three clusters of men and women inferred at all three time-points (baseline, 4 weeks and 8 weeks) of the trial. The pie charts at 4 weeks and 8 weeks illustrate the patient composition based on their cluster assignment at baseline. Patient stratification can then be used to form the factor graph as shown in Fig. (b). The graph is bipartite, with patient stratification (one set of nodes) and factor functions (the other set of nodes) that capture relationships between the symptom severity associated with the stratification and other data.

4. Using Factor Graphs to Model Symptom Dynamics

Patient stratification (clusters) was obtained using only the symptom severity scores, which allowed us to establish broader ranges of symptom severity in the trial. This stratification will now be used to establish the symptom dynamics in response to the antidepressants. *Symptom dynamics* in this work is defined as the likely changes in a patient's symptoms and associated clinical outcomes during the various stages of the trial (e.g., *response* at 4 weeks or 8 weeks) while he or she is treated with antidepressants.

Factor graph from patient stratification: The factor graph(for men and women separately) is a bipartite graph $\mathcal{G} = (\mathcal{V}, \mathcal{F})$ (we created separate factor graphs for men and women). The graph has three layers at each time point as illustrated in Fig. 3(b); the *clinical observation layer* where the clinician observes the clinical outcome based on symptom severity, *patient symptom response layer* that keeps track of changes in symptoms and the *patient stratification layer* to illustrate what cluster a patient's symptom score belongs to. Each layer is associated one variable node $\in \mathcal{V}$ such as $\mathcal{O}$ (distribution of patients who demonstrate response (R) vs no response (NR)), $\mathcal{X}$ (symptom measure at each time point), $\mathcal{C}$ (patient stratification at each time point) and one associated factor node $\in \mathcal{F}$ such as a decision rule to determine if a patient has demonstrated response (50% reduction in symptom from baseline) for random variable $\mathcal{O}$, a transition probability matrix for symptom severity between two time points for random variable $\mathcal{X}$, and what cluster $\mathcal{C}$ the patient belongs to based on his or her current symptom severity score $\mathcal{X}$. The graph can be evaluated at each time point of the trial $t \in T$ starting from baseline (t) to 4 weeks ($t+1$) to 8 weeks ($t+2$) and so on.

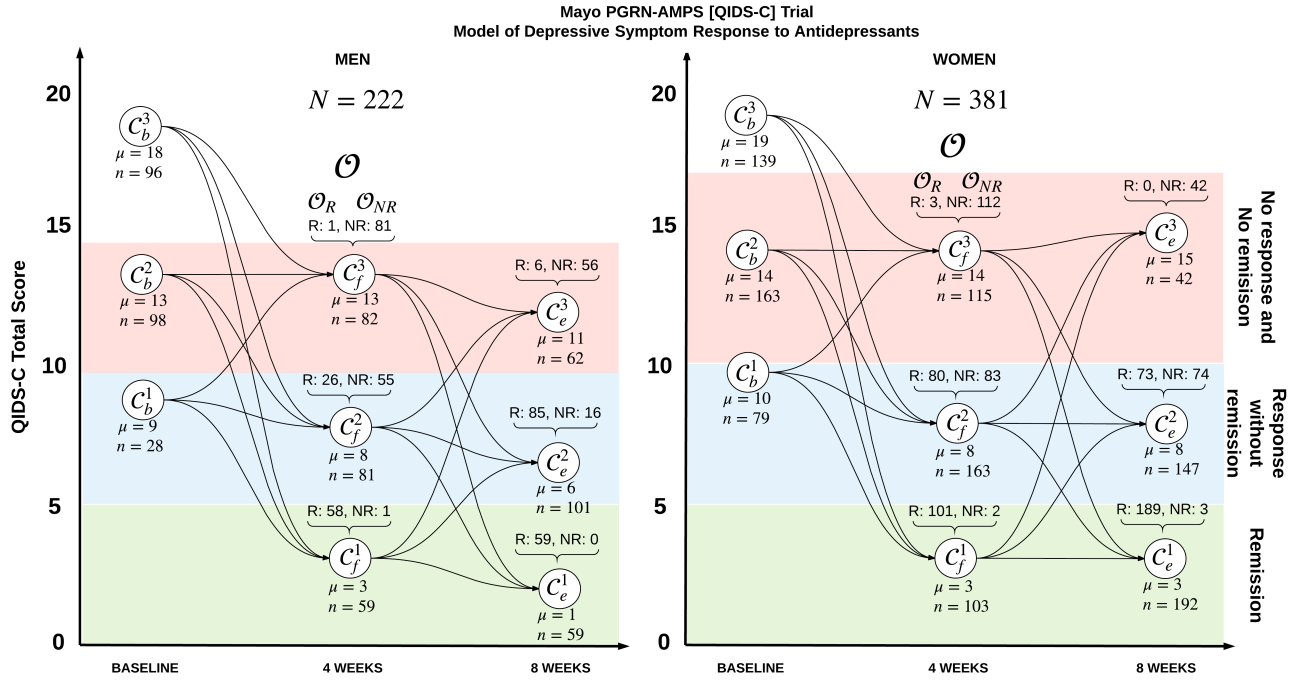
The maximum likelihood symptom response to antidepressants: We use the forward algorithm [16], [17] to identify the most likely *forward* transitions a patient starting in any baseline cluster will make between clusters (*hidden states* –$\mathcal{C}$) of the trial, and also what the associated clinical outcomes will be during the transitions (*observed states* –$\mathcal{O}$). During transitions between the clusters, the clinician/psychiatrist assessing the patient observes the clinical outcome $\mathcal{O} = \{\mathcal{O}_R, \mathcal{O}_{NR}\}$, which is that the patient has demonstrated either $\mathcal{O}_R$–*response*, or $\mathcal{O}_{NR}$– *no-response*. For both men and women, the graph with the number of patients (n), forward transitions, and observed outcomes $\mathcal{O} = \{\mathcal{O}_R, \mathcal{O}_{NR}\}$ in each cluster are illustrated in Fig. 4(a), which is similar in construct to a hidden Markov model (HMM). Now, the symptom dynamics for any patient starting in any of the clusters at baseline can be solved recursively using the *forward algorithm* which is described as,

$$P_{\mathcal{O}}(\mathcal{C}_t) = \sum_{t \in T} p(\mathcal{O}|\mathcal{C}_t) P_{\mathcal{O}}(\mathcal{C}_{t-1}) p(\mathcal{C}_{t-1} \rightarrow \mathcal{C}_t), \quad (2)$$

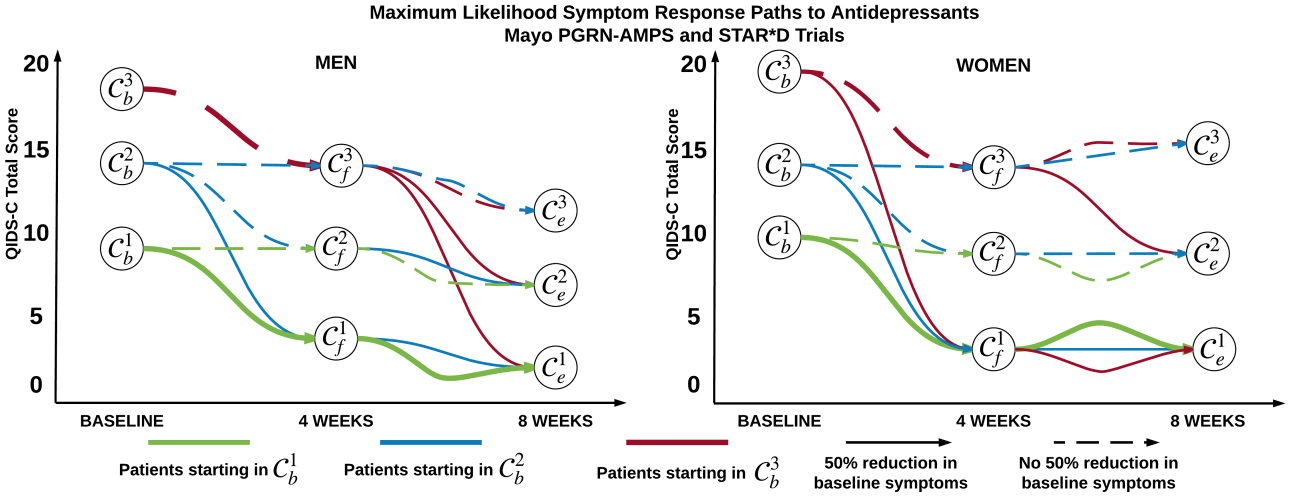
where $p(\mathcal{O}|\mathcal{C}_t)$ is the probability of the observation (response or no-response) in a current state, $p(\mathcal{C}_{t-1} \rightarrow \mathcal{C}_t)$ is the probability of a transition from a state of a previous time-point to a state of the current time-point (e.g., $\mathcal{C}_b^1 \rightarrow \mathcal{C}_f^2$), and $P_{\mathcal{O}}(\mathcal{C}_{t-1})$ is the path probability for a given set of observations $\mathcal{O}$ seen until $\mathcal{C}_{t-1}$.

Note that, *the reduction from the factor graph to the HMM did not simplify the complexity of solving the forward algorithm, but rather allowed us to explain the symptom dynamics not only as a function of how symptoms change, but also with the changes in symptoms, what are the associated clinical outcomes during various time-points of the trial.*

Results: For every cluster starting at the baseline, the path probabilities for all possible combinations of paths and observations to get to a state at 8 weeks were computed using Equation 2. For each starting cluster at baseline, we



(a) Patient stratification and associated composition of clusters by the associated clinical outcomes (R = *response*, NR = *no response*).



(b) Maximum likelihood symptom response paths to antidepressants.

Figure 4: Symptom dynamics using HMM. Fig. (a) illustrates the HMM of the symptom dynamics in men and women; Fig. (b) illustrates the inferred most likely symptom dynamics based on the cluster in which each patient starts in the trial. Thicker lines indicate a larger proportion of patients from the originating cluster taking a particular path.

chose the path that had the highest probability of ending at each of the clusters at 8 weeks and also at least 20% of the cohort in the path, so that we do not choose paths of symptom changes that only a few patients take, which renders lesser statistical power. For example, for patients starting in $\mathcal{C}_b^1$, we inferred the best path to $\{\mathcal{C}_e^1, \mathcal{C}_e^2, \mathcal{C}_e^3\}$ and the associated outcomes that maximize the path probabilities as shown in Fig. 4(b). The interpretation of the results are as follows.

Men:

1) A majority of the men starting in $\mathcal{C}_b^1$ will most likely *respond* at 4 weeks, and they will most-likely go on to achieve *remission* (and demonstrate *response*) at 8 weeks. However, if they do not *respond* at 4 weeks, they will likely not *respond* at 8 weeks either.
2) A majority of the men starting in $\mathcal{C}_b^2$ at baseline will most likely not *respond* at 4 weeks, will also not *respond* at 8 weeks. However, if they *respond* at 4 weeks, they

will likely achieve *remission* at 8 weeks.

3) Men starting in $\mathcal{C}_b^3$ at baseline are not likely to *respond* at 4 weeks, but two-thirds of them respond by the end of 8 weeks.

Women:

1) Regardless of where women start at baseline, if they *respond* by the end of 4 weeks, they are almost certain to achieve remission at 8 weeks.
2) If women starting in $\mathcal{C}_b^1$ do not *respond* at 4 weeks, they will likely not *respond* at 8 weeks either.
3) A majority of the women starting in $\mathcal{C}_b^2$ at baseline will most likely not *respond* at 4 weeks, will also not *respond* at 8 weeks.
4) Women who do not *respond* at 4 weeks are more likely to *respond* at 8 weeks and have moderate symptoms ($\mathcal{C}_e^2$), than have relatively more severe symptoms ($\mathcal{C}_f^3$).

To the best of our knowledge, this is the first work that formally establishes the symptom dynamics in major depressive disorder. We will discuss the clinical implications of these paths in Sec. 6.

5. Using Baseline Data to Predict Clinical Outcomes

Having identified the most likely symptom responses to antidepressants, we focus on predicting clinical outcomes using only baseline data. The developed factor graph for modeling symptom dynamics allows for integration of metabolomics data for each of the clusters through use of more complicated factor functions for predicting which state the patient could fall into after 8 weeks of treatment. Now, we also know that in general, the inferred clusters of patients map well to the cohorts classified according to known psychiatric definitions, such as those of responders vs non-responders ($\mathcal{C}_e^3$ vs. the rest) and remitters vs. non-remitters ($\mathcal{C}_e^1$ vs. the rest). The metabolomics data was collected from a smaller cohort of the Mayo PGRN-AMPS trial, which meant that for some clusters we only had less than 10 patients, rendering limited statistical power. Hence we decided to train two binary prediction models to predict 1) whether a patient will demonstrate either response or no response, and 2) whether remission or no remission will be achieved at the end of 8 weeks.

Recent work proposed a prediction model that uses elastic-net regularization for feature selection and a gradient boosting machine (GBM) for classification, but only using baseline social, demographic, and clinical data from the STAR*D trial [22]. While their prediction accuracies were better than chance, the authors acknowledge the limitations of their work, which suggests that it might be worthwhile to study whether the addition of baseline biological measures together with the social, demographic, and clinical data would increase the predictability of the clinical outcomes. With access to metabolomics data on a smaller cohort of the Mayo PGRN-AMPS trial, we set out to test whether a similar combination of feature selection and a classifier could improve the predictability of clinical outcomes.

Feature selection and choice of classifiers: Three classes of classifiers are used in this work, including kernel, linear, and ensemble methods. We used support vector machines with linear kernels (SVMLinear) and support vector machines that use radial-basis kernels (SVM-RBF) as kernel methods [23], a generalized linear model (GLM) as a linear method [24] and gradient-boosting machines (GBM) as an ensemble method [25]. As authors of these methods have indicated, each of these broader types has their own merits, mathematical nuances, and complexities, and all of them have been used in other classification applications, such as in Kaggle [26].

In addition to elastic-net regularization, recursive feature elimination (i.e., a wrapper method) was also used for the GLM and GBM classifiers that made it possible to estimate the model performance by not only optimizing the parameters of the model, but also by searching for the right set of predictor variables. Based on our datasets, the prediction performance did not significantly vary with or without the use of any of the feature selection methods; the prediction accuracy remained within 4%.

Furthermore, to address the bias-variance tradeoffs, we performed tenfold cross-validation with 3 repeats and an expansive grid-search for the parameter space for the classifiers in order to train the classifiers on 80% of the overall data; the remaining 20% was used for testing the trained models. A combination of the overall accuracy i.e., the fraction of labels that were correctly predicted) and the area under the receiver operating characteristic curve (AUC) metric was used to choose the training model [27]. We also computed the sensitivity and specificity of the prediction models.

Training with and without biological measures: In order to quantitatively assess the benefit of adding biological measures to predict outcomes, we trained classifiers using 1) baseline clinical data that included only social and demographic data, $X = [S : C]$; and 2) all baseline data (including metabolomics data), $X = [S : C : B]$. Metabolites whose baseline concentrations were correlated with the symptom severity at 8 weeks were chosen, and then normalized along with clinical and demographic data in order to train the chosen supervised learning methods. Since the clinical data were assessed on all the Mayo PGRN-AMPS trial by the same four physicians, we believe that the proposed mechanism of combining clinical data with biological measures will not result in the loss of any underlying characteristics of either type of the data. Several other researchers have proposed the combination other modalities of biological data, but not together with clinical data that includes a significant amount of a patient responses to symptom questionnaire [12]–[14]. The limitations in those proposals are that either biological measures of the same type are fused with data from various sources, or different biological measures are combined based on existing biological knowledge in the context of the disease. Therefore, to the best of our best knowledge, this is the first time a more accurate biological measure in the context of psychiatry has been integrated for analyses with the clinical measures that comprise demographic data and patient-provided responses to symptom questionnaire. For

TABLE 2: Clinical outcome prediction performance for men in Mayo Clinic PGRN-AMPS trial. Expansion of abbreviations of the top predictors are as follows, ATOCO is (+)-alpha-Tocopherol; URIC is Uric acid; QIDS-1 is sleep-onset insomnia [19]; KYN is Kynurenine; 3OHKY is 3-Hydroxykynurenine; AMTRP is Alpha-methyltryptophan; I3PA is Indole-3-propionic acid; GTOCO3 is (+)-gamma-Tocopherol (redox state #3); 5HT is serotonin.

RESPONSE									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors: GLM
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	ATOCO
Accuracy	28.2	32	52	40	48	48	64	48	URIC
Sensitivity	0	16.67	16.67	33.33	33.33	33.33	50	33.33	QIDS-1
Specificity	53.5	46.15	84.62	46.15	61.54	61.54	61.54	61.54	KYN
AUC	0.64	0.60	0.63	0.54	0.53	0.53	0.68	0.5	3OHKY
REMISSION									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors: SVM-Linear
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	AMTRP
Accuracy	28	44	44	48	64	68	64	45.65	I3PA
Sensitivity	38.46	38	53.85	46.15	76.52	76	76.92	65.22	Drug dosage
Specificity	16.67	50	33.33	50	50	50	50	26.09	GTOCO3
AUC	0.8	0.6	0.67	0.6	0.76	0.78	0.62	0.6	5HT

TABLE 3: Clinical outcome prediction performance for women in Mayo Clinic PGRN-AMPS trial. Expansion of abbreviations of the top predictors are as follows 5HT is Serotonin; MHPG is Methoxy-Hydroxyphenly Glycol; MET is Methionine; QIDS-13 is involvement [19]; HGA is Homogentisic Acid; 3OHKY is 3-Hydroxykynurenine; PARAXAN is 1,7-diMethylxanthine.

RESPONSE									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors: SVM-Linear & GLM
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	Seasonal Pattern
Accuracy	52.08	52.08	54.17	50	41.3	72.33	64.58	41.67	5HT
Sensitivity	18.18	18.18	27.27	18.18	34.78	18.18	36.36	0	MHPG
Specificity	80.72	80.76	76.92	76.9	47.83	92.83	88.46	76	MET
AUC	0.60	0.59	0.63	0.63	0.69	0.74	0.68	0.49	QIDS-13
REMISSION									
Data	Clinical Data Only				Clinical and Metabolomics Data				Top Predictors: SVM-Linear
Model	SVM-RBF	SVM-Linear	GLM	GBM	SVM-RBF	SVM-Linear	GLM	GBM	5HT
Accuracy	34.78	50	45.65	36.96	41.3	54.33	52.17	45.65	HGA
Sensitivity	26.09	65.22	56.52	47.83	34.78	56.52	76.92	65.22	3OHKY
Specificity	43.48	34.78	34.78	26.09	47.83	52.17	50	26.09	Seasonal Pattern
AUC	0.64	0.48	0.42	0.58	0.56	0.53	0.53	0.47	PARAXAN

all the classifiers, we compared the AUC, in addition to the generalized prediction accuracies to see if the same models predictive ability improved with the addition of metabolomics data. Further, if the predictability improved, we extracted the top five predictors of the model that provided the best balance of accuracy and AUC to see if the top predictors were dominated by the metabolomics.

Results: As shown in Tables 2 and 3, for both men and women and for both outcomes *response* and *remission*, at least 3 of the 4 methods showed an improvement of the AUC and the corresponding overall accuracy, i.e., the proportion of samples correctly predicted with the addition of the metabolomics data. The highlighted columns in Tables. 2 and 3 indicate the best-performing models with the metabolomics data included; 4 out of the 5 predictors are metabolites, indicating that their addition to the prediction model likely explains the increase in the predictability of the outcomes. Top predictor metabolites were also different in men and women, pointing to a likely different biological mechanism in how men and women respond to the same antidepressant. Further, many of the top predictor metabolites identified in this work are known to be correlated with mood in behavioral sciences, which leads to additional promising implications as discussed next.

6. From Analytics to the Clinical Implications

Patient stratification conforming to known psychiatric definitions of outcomes and replicated in both the Mayo PGRN-AMPS and STAR*D trial provided a first show of confidence in modeling symptom responses to citalopram/escitalopram (antidepressant) treatment in depressed patients. The factor graph model was able to show potentially important clinical differences between men and women in depressive symptom behavior over time under antidepressant treatment. The unique patterns of symptom dynamics we found in men and women will add to the psychiatric community's increasing evidence of sex differences in terms of responses to treatment in patients with major depression. Our focus on remission, which is defined clinically as a relative absence of depressive symptoms, was motivated by the fact that failure to achieve remission is associated with ongoing difficulties with psychosocial functioning owing to residual depressive symptoms, and higher odds of full depressive relapses, even for patients who achieve a positive response. For women, achieving response to antidepressant treatment at 4 weeks strongly predicted remission at 8 weeks regardless of baseline depressive symptom severity. For men, the same was also true, but only for those in the low ($\mathcal{C}_b^1$) and moderate ($\mathcal{C}_b^2$) symptom clusters at baseline. In those groups, failure to

achieve response at 4 weeks was highly predictive of a lack of remission at 8 weeks, with the only exception being for men in the most severe symptom cluster at baseline ($\mathcal{C}_b^3$). In that group of men with more severe depression at baseline, the odds of response at 4 weeks were low; however, two-thirds of these individuals either responded or remitted by week 8. In general, our results support clinical recommendations for examining depressive symptoms after 4-6 weeks of treatment before judging the clinical effects of antidepressant treatment. Our results suggest that ascertaining clinical effects of an antidepressant at 4 weeks may be especially reasonable for women, regardless of their baseline symptom severity, and for men with milder-to-moderate depressive symptoms at baseline. For men with more severe depression at the start of treatment, the time window may need to be extended beyond 4 weeks before the full effects of treatment with antidepressant at a given dose can be judged.

On two other fronts, this preliminary study of evaluating the predictability of outcomes when metabolomics data are added as a biological measure has been promising. First, there was an overall improvement in the accuracy of predictions, and second, known metabolites were found to be among the top predictors of the outcomes. Specifically, for decades the treatment of MDD has focused on biogenic amine neurotransmitter pathways, i.e., the synthesis and metabolism of catecholamines such as norepinephrine and indoleamines such as serotonin [28]. Furthermore, existing body of knowledge fits well with the findings of our study; note that the metabolites listed in Tables 2 and 3 include serotonin (5HT) itself as well as two metabolites from the competing tryptophan metabolism pathway (KYN and 3OHKYN) and the major catecholamine metabolite (MHPG), which are known to play a role in behavioral sciences.

7. Conclusions and Future Work

Data-driven analytics using factor graphs and a variety of learning methods based on data from two of the largest trials in major depressive disorder established consistent stratification of patients and establish symptom dynamics in depressed patients treated with antidepressants. These findings agreed with existing definitions of psychiatric outcomes and results were replicated in two of the largest clinical trials in USA. Furthermore, addition of biological measures such as the metabolomics to baseline social, demographic and clinical data significantly improved predictability in clinical outcomes at 8 weeks. Top predictors were significantly different in men and women, and were known to be coming from metabolomism pathways known to the psychiatric community. These results also suggest several questions that we will strive to answer in the course of our future work. First, are the clusters of patients in anyways associated with any of the social, demographic and clinical factors? Second, are such associations related to any baseline metabolomic concentrations that could improve the predictability of outcomes? Finally, we could as whether the baseline metabolomics data alone might better predict clinical outcomes?

8. Tool Implementation and Availability

The analyses' workflow was developed in R, version 3.2.2. The workflow has been tested on and is compatible with Linux (Ubuntu 14 and later), OS X (Yosemite and later) and Windows (Windows 7 and later) operating systems. Further, in addition to being compatible with Intel's x86 architectures, this workflow is also compatible with the Power Architecture, and was tested on the IBM POWER8 processors, establishing that our tool is agnostic with respect to common, state-of-the-art high-performance computing platforms. The platforms and operating-system agnostic characteristics of our workflow means that it can easily be integrated into larger metabolomics studies, and other -omics data such as the genomics (genotype), both on stand-alone machines and in the cloud. We intend to making the workflow and sample test data publicly available on the acceptance of this paper (the data from the STAR*D trial is already publicly available).

9. Related Work

Patient stratification in itself is not a new concept in medicine or clinical research, however, given the inter-patient variability in the symptoms assessed in patients with MDD, there has been no clear definition of patient stratification in terms of their symptom severity. Since studying symptom dynamics needs the general grouping of patients in a trial, without patient stratification, there are also no models to study the symptom dynamics in patients with MDD treated with antidepressants. Hence, to the best of our knowledge, this work presents a new contribution to the existing body of literature in psychiatry, that demonstrated shown cross-trial validation.

Integration of multi-modal biological data has been proposed in the context of breast cancer, diabetes, glioblastoma and other diseases [12]–[15]. For example, combining imaging data with gene expression and patient demographics data to better predict the prevalence of cancer [15]. These aforementioned works have looked at identifying unique biological signatures or biomarkers from each of the types of static data, and then building a function that linearly or non-linearly combines the data into a rule-based decision system. However, in diseases such as MDD where symptoms change with time due to the therapeutic design, then models need to incorporate the time dimension as well. Furthermore, in diseases like MDD where symptom assessment is subject to differences in inter-patient variability, models need to be amenable to rely more on the accurate biological measures while still considering effects of symptom response. Addressing these needs, the stratification used clinician's ratings of total symptom severity and minimized the effects of inter-patient variability in assessing the response to symptom questionnaire. This approach proved to be valuable because of the ability to replicate the stratification and symptom dynamics in two of the largest clinical trials in USA. While our current approach to integrate the biological measure such as metabolomics with clinical measures is simple due to

limited data, the predictability and the biological significance of predictor metabolites in the context of psychiatry is promising and encourages us to collect and analyze more data to further the understanding of what mechanisms drives response to antidepressants.

10. Acknowledgments

This material is based upon work partially supported by a Mayo Clinic and Illinois Alliance Fellowship for Technology-Based Healthcare Research; a CompGen Fellowship; an IBM Faculty Award; National Science Foundation (NSF) under grants CNS 13-37732, CNS 16-24790 and CNS 16-24615; National Institutes of Health (NIH) under grants U19 GM61388, RO1 GM28157, R24 GM078233 and RC2 GM092729; and The Mayo Clinic Center for Individualized Medicine. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the NSF and NIH. We thank Jenny Applequist for her help in preparing the manuscript.

References

[1] N. Olchanski, M. M. Myers, M. Halseth, P. L. Cyr, L. Bockstedt, T. F. Goss, and R. H. Howland, "The economic burden of treatment-resistant depression," *Clinical therapeutics*, vol. 35, no. 4, pp. 512–522, 2013.

[2] K. Martinowich, D. Jimenez, C. Zarate, and H. Manji, "Rapid antidepressant effects: moving right along," *Molecular psychiatry*, vol. 18, no. 8, pp. 856–863, 2013.

[3] R. C. Kessler, H. S. Akiskal, M. Ames, H. Birnbaum, P. Greenberg, R. Jin, K. R. Merikangas, G. E. Simon, P. S. Wang *et al.*, "Prevalence and effects of mood disorders on work performance in a nationally representative sample of us workers," *American Journal of Psychiatry*, 2006.

[4] M. H. Trivedi, M. Fava, S. R. Wisniewski, M. E. Thase, F. Quitkin, D. Warden, L. Ritz, A. A. Nierenberg, B. D. Lebowitz, M. M. Biggs *et al.*, "Medication augmentation after the failure of ssris for depression," *New England Journal of Medicine*, pp. 1243–1252, 2006.

[5] F. A. Jain, A. M. Hunter, J. O. Brooks, and A. F. Leuchter, "Predictive socioeconomic and clinical profiles of antidepressant response and remission," *Depression and anxiety*, pp. 624–630, 2013.

[6] R. Hirschfeld, J. M. Russell, P. L. Delgado, J. Fawcett, R. A. Friedman, W. M. Harrison, L. M. Koran, I. W. Miller, M. E. Thase, R. H. Howland *et al.*, "Predictors of response to acute treatment of chronic and double depression with sertraline or imipramine." *The Journal of clinical psychiatry*, 1998.

[7] R. M. Bagby, A. G. Ryder, and C. Cristi, "Psychosocial and clinical predictors of response to pharmacotherapy for depression," *Journal of psychiatry & neuroscience: JPN*, p. 250, 2002.

[8] A. C. Altamura, C. Montresor, D. Salvadori, and E. Mundo, "Does comorbid subthreshold anxiety affect clinical presentation and treatment response in depression? a preliminary 12-month naturalistic study," *International Journal of Neuropsychopharmacology*, pp. 481–487, 2004.

[9] R. Iniesta, K. Malki, W. Maier, M. Rietschel, O. Mors, J. Hauser, N. Henigsberg, M. Z. Dernovsek, D. Souery, D. Stahl *et al.*, "Combining clinical variables to optimize prediction of antidepressant treatment outcomes," *Journal of psychiatric research*, pp. 94–102, 2016.

[10] D. A. Mrazek, J. M. Biernacka, D. J. O'kane, J. L. Black, J. M. Cunningham, M. S. Drews, K. A. Snyder, S. R. Stevens, A. J. Rush, and R. M. Weinshilboum, "Cyp2c19 variation and citalopram response," *Pharmacogenetics and genomics*, 2011.

[11] M. H. Trivedi, A. J. Rush, S. R. Wisniewski, A. A. Nierenberg, D. Warden, L. Ritz, G. Norquist, R. H. Howland, B. Lebowitz, P. J. McGrath *et al.*, "Evaluation of outcomes with citalopram for depression using measurement-based care in star* d: implications for clinical practice," *American journal of Psychiatry*, pp. 28–40, 2006.

[12] D. Lahat, T. Adali, and C. Jutten, "Multimodal data fusion: an overview of methods, challenges, and prospects," *Proceedings of the IEEE*, vol. 103, no. 9, pp. 1449–1477, 2015.

[13] B. Ray, M. Henaff, S. Ma, E. Efstathiadis, E. R. Peskin, M. Picone, T. Poli, C. F. Aliferis, and A. Statnikov, "Information content and analysis methods for multi-modal high-throughput biomedical data," *Scientific reports*, vol. 4, p. 4411, 2014.

[14] P.-Y. Wu, C.-W. Cheng, C. D. Kaddi, J. Venugopalan, R. Hoffman, and M. D. Wang, "–omic and electronic health record big data analytics for precision medicine," *IEEE Transactions on Biomedical Engineering*, vol. 64, no. 2, pp. 263–273, 2017.

[15] J. Kong, L. A. Cooper, F. Wang, D. A. Gutman, J. Gao, C. Chisolm, A. Sharma, T. Pan, E. G. Van Meir, T. M. Kurc *et al.*, "Integrative, multimodal analysis of glioblastoma using tcga molecular data, pathology images, and clinical outcomes," *IEEE Transactions on Biomedical Engineering*, vol. 58, no. 12, pp. 3469–3474, 2011.

[16] B. J. Frey, F. R. Kschischang, H.-A. Loeliger, and N. Wiberg, "Factor graphs and algorithms."

[17] H.-A. Loeliger, "An introduction to factor graphs," *IEEE Signal Processing Magazine*, vol. 21, no. 1, pp. 28–41, 2004.

[18] A. P. Athreya, K. Y. Ngiam, Z. Luo, E. S. Tai, Z. Kalbarczyk, and R. K. Iyer, "Towards longitudinal analysis of a population's electronic health records using factor graphs," in *Proceedings of the 3rd IEEE/ACM International Conference on Big Data Computing, Applications and Technologies*. ACM, 2016, pp. 79–86.

[19] A. J. Rush, M. H. Trivedi, H. M. Ibrahim, T. J. Carmody, B. Arnow, D. N. Klein, J. C. Markowitz, P. T. Ninan, S. Kornstein, R. Manber *et al.*, "The 16-item quick inventory of depressive symptomatology (qids), clinician rating (qids-c), and self-report (qids-sr): a psychometric evaluation in patients with chronic major depression," *Biological psychiatry*, pp. 573–583, 2003.

[20] M. Piccinelli and G. Wilkinson, "Gender differences in depression critical review," *The British Journal of Psychiatry*, vol. 177, no. 6, pp. 486–492, 2000.

[21] J. D. Banfield and A. E. Raftery, "Model-based gaussian and non-gaussian clustering," *Biometrics*, pp. 803–821, 1993.

[22] A. M. Chekroud, R. J. Zotti, Z. Shehzad, R. Gueorguieva, M. K. Johnson, M. H. Trivedi, T. D. Cannon, J. H. Krystal, and P. R. Corlett, "Cross-trial prediction of treatment outcome in depression: a machine learning approach," *The Lancet Psychiatry*, pp. 243–250, 2016.

[23] B. Scholkopf and A. J. Smola, *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.

[24] J. Friedman, T. Hastie, and R. Tibshirani, "Regularization paths for generalized linear models via coordinate descent," *Journal of statistical software*, vol. 33, no. 1, p. 1, 2010.

[25] J. H. Friedman, "Stochastic gradient boosting," *Computational Statistics & Data Analysis*, vol. 38, no. 4, pp. 367–378, 2002.

[26] J. Friedman, T. Hastie, and R. Tibshirani, *The elements of statistical learning*. Springer series in statistics Springer, Berlin, 2001, vol. 1.

[27] J. Demšar, "Statistical comparisons of classifiers over multiple data sets," *Journal of Machine learning research*, vol. 7, no. Jan, pp. 1–30, 2006.

[28] J. J. Schildkratu, "Neuropsychopharmacology and the affective disorders," 1969.