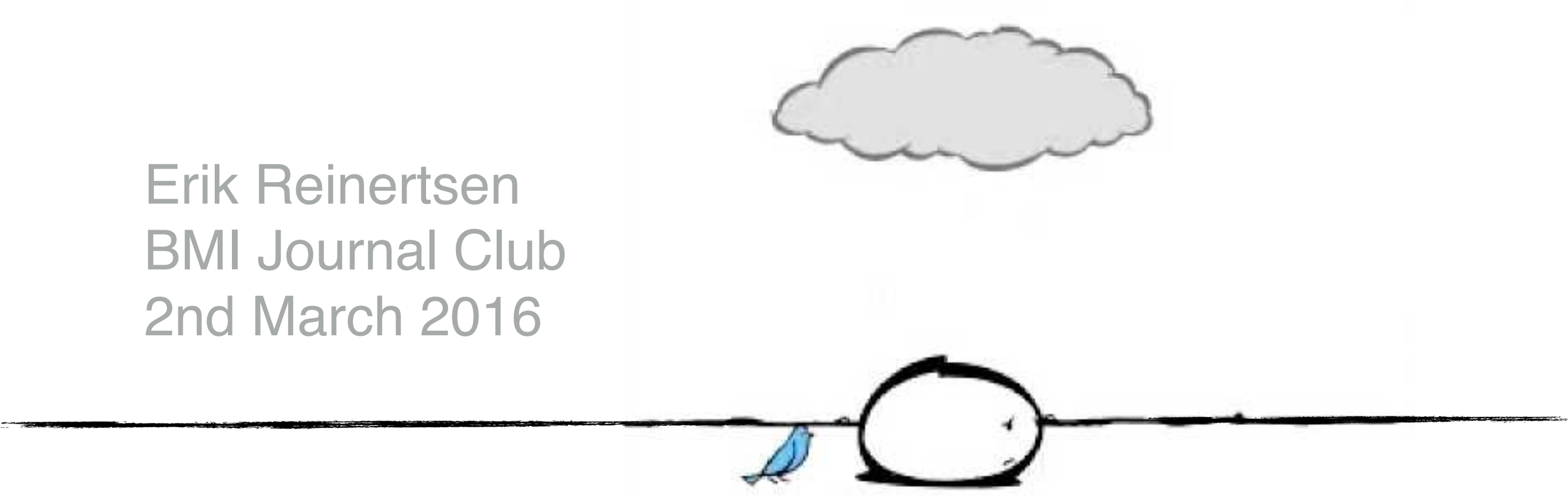
Data mining methods for sub-typing major depressive disorder

Erik Reinertsen
BMI Journal Club
2nd March 2016



Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Van Loo, H. M. et al. Major depressive disorder subtypes to predict long-term course. Depress. Anxiety (2014).

Problem: variation in course of MDD is not strongly predicted by existing subtype distinctions.

Novelty: use data mining techniques to find subtypes *(based on index episode symptoms)* that predict subsequent MDD course.

Results: found 3 clusters of high, intermediate, or low predicted outcome scores.

Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

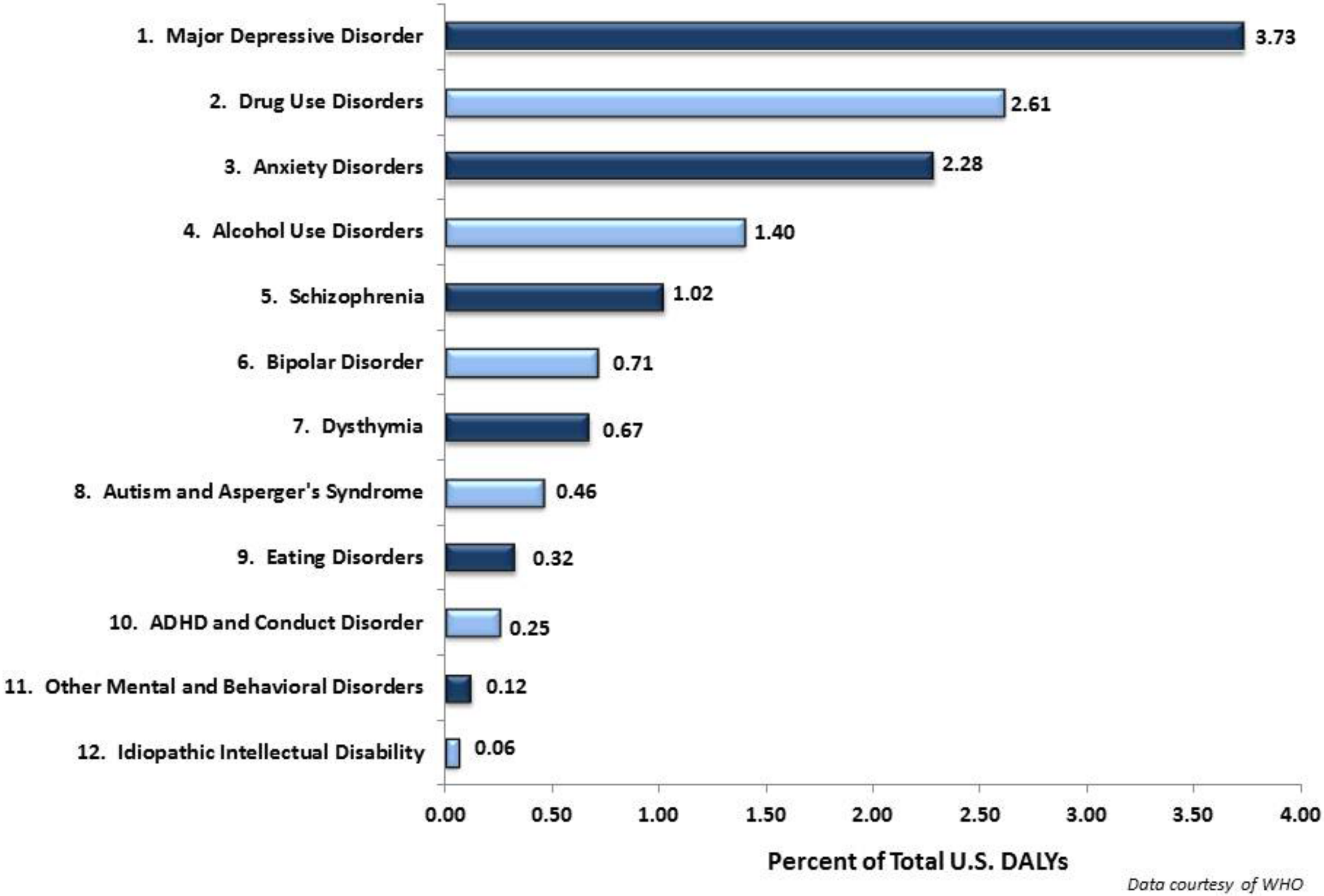
Major Depressive Disorder

…is characterized by a history of 1+ major depressive episodes and no history of mania or hypomania.

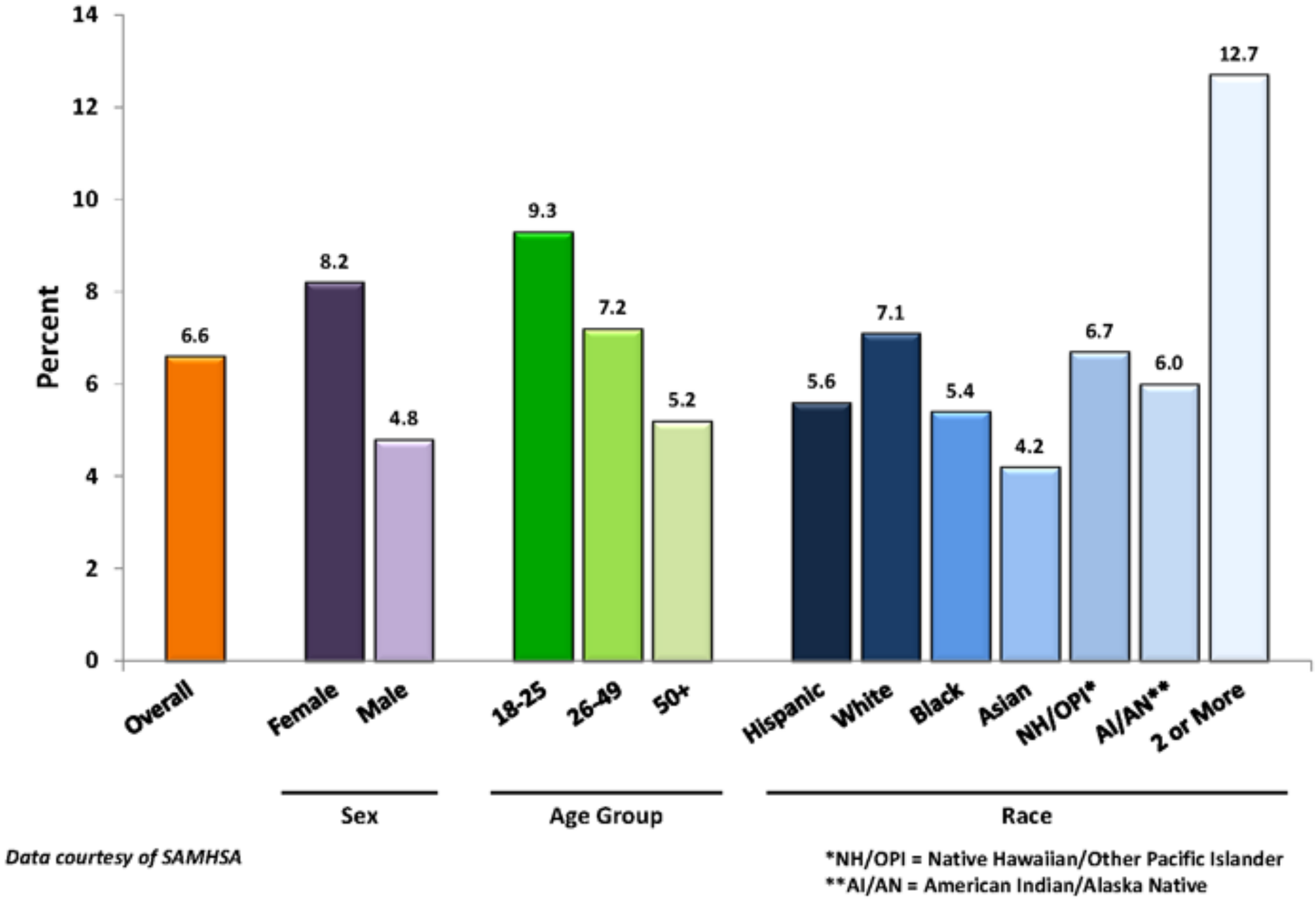
A major depressive episode manifests with 5+ of the following 9 symptoms for at least 2 consecutive weeks; at least one symptom must be either depressed mood or loss of interest or pleasure…

In addition, the symptoms cause significant distress or psychosocial impairment, and are not the direct result of a substance or general medical condition.

U.S. DALYs for Mental and Behavioral Disorders as a Percent of Total U.S. DALYs (2010)

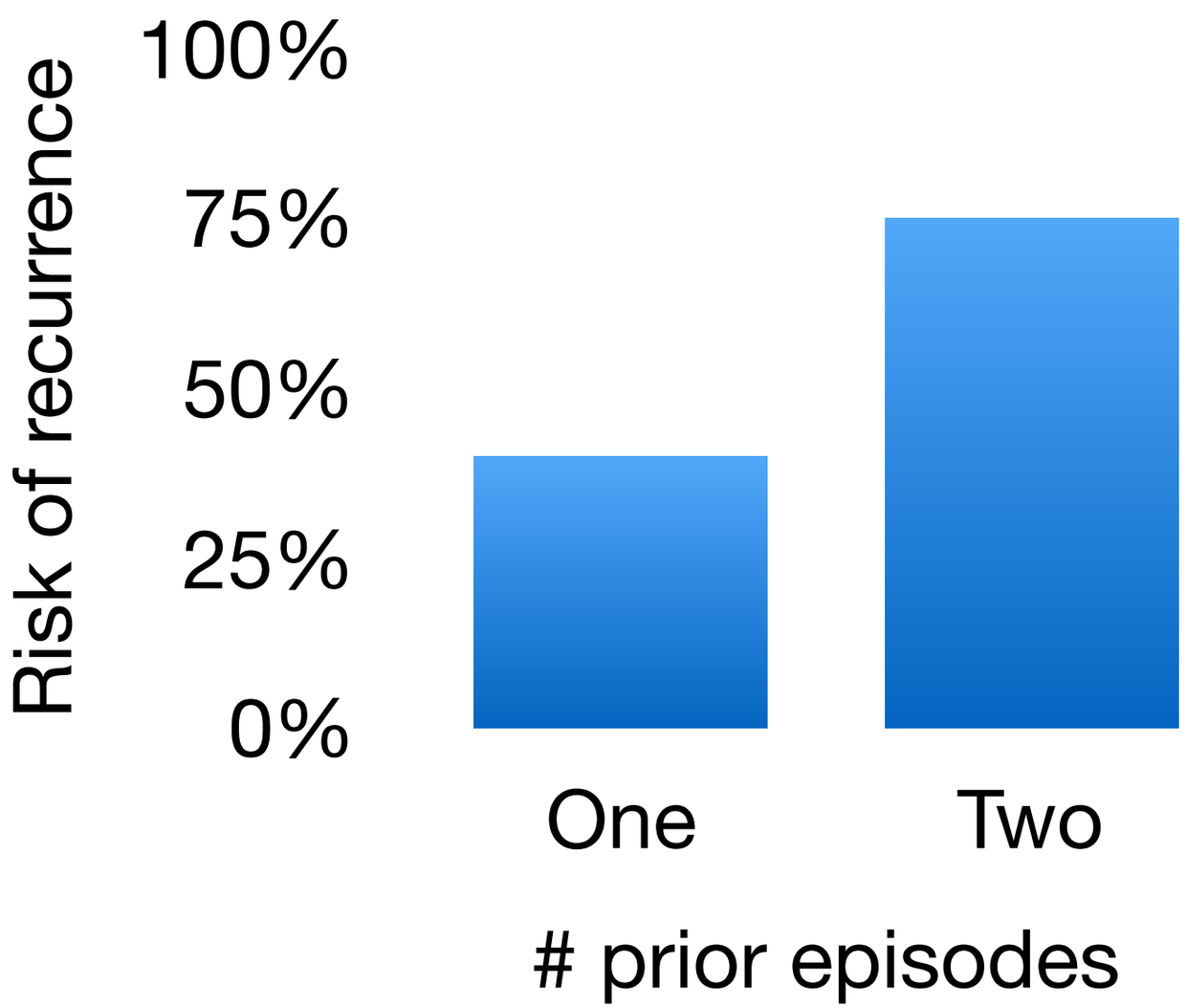


Data courtesy of WHO



Data from 2014 National Survey on Drug Use and Health (NSDUH)

MDD is highly recurrent



Solomon, D. A., et al. Am J Psych (2001)

MDD varies substantially in treatment response and illness course

Symptomatic diversity hinders research of mechanisms and treatment

"Causal" e.g. postnatal; speculative.

Symptom-driven e.g. atypical or melancholic depression; DSM-5, but lacks clinical utility.

Empirical e.g. cluster, factor, or latent class analysis but lacks evidence![1]

1. van Loo et al. BMC Med (2012)

Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Data

WHO World Mental Health surveys

16 countries (high, middle, & lower income)

93,167 adult participants

8,261 people met lifetime DSM-IV criteria for MDD

Average response rate: 74%

Questions asked of people with lifetime MDD:

Age of onset

Role of stressful life experience in first depressive episode

DSM-IV Criterion A-D symptoms

ICD-10 severity specifiers

Four outcomes of this study

1. # years since onset of episode lasting 2+ wks*
2. # years since onset of episode lasting *most days* of the year*
3. Overnight hospitalization from depression (and if so, age)
4. Disability because of depression

*transformed to continuous data by dividing by # years between age at interview and AOO+1

Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Three technical methods

1. Random forest
2. Lasso GLM
3. *k*-means clustering

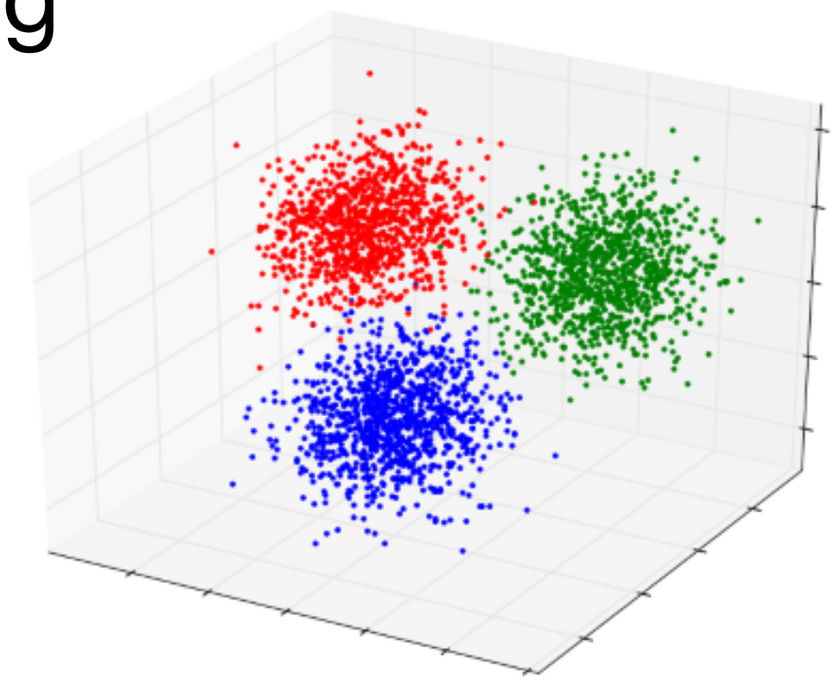
Random forest: to find important predictor interactions

Lasso GLM: to predict outcomes from best predictors

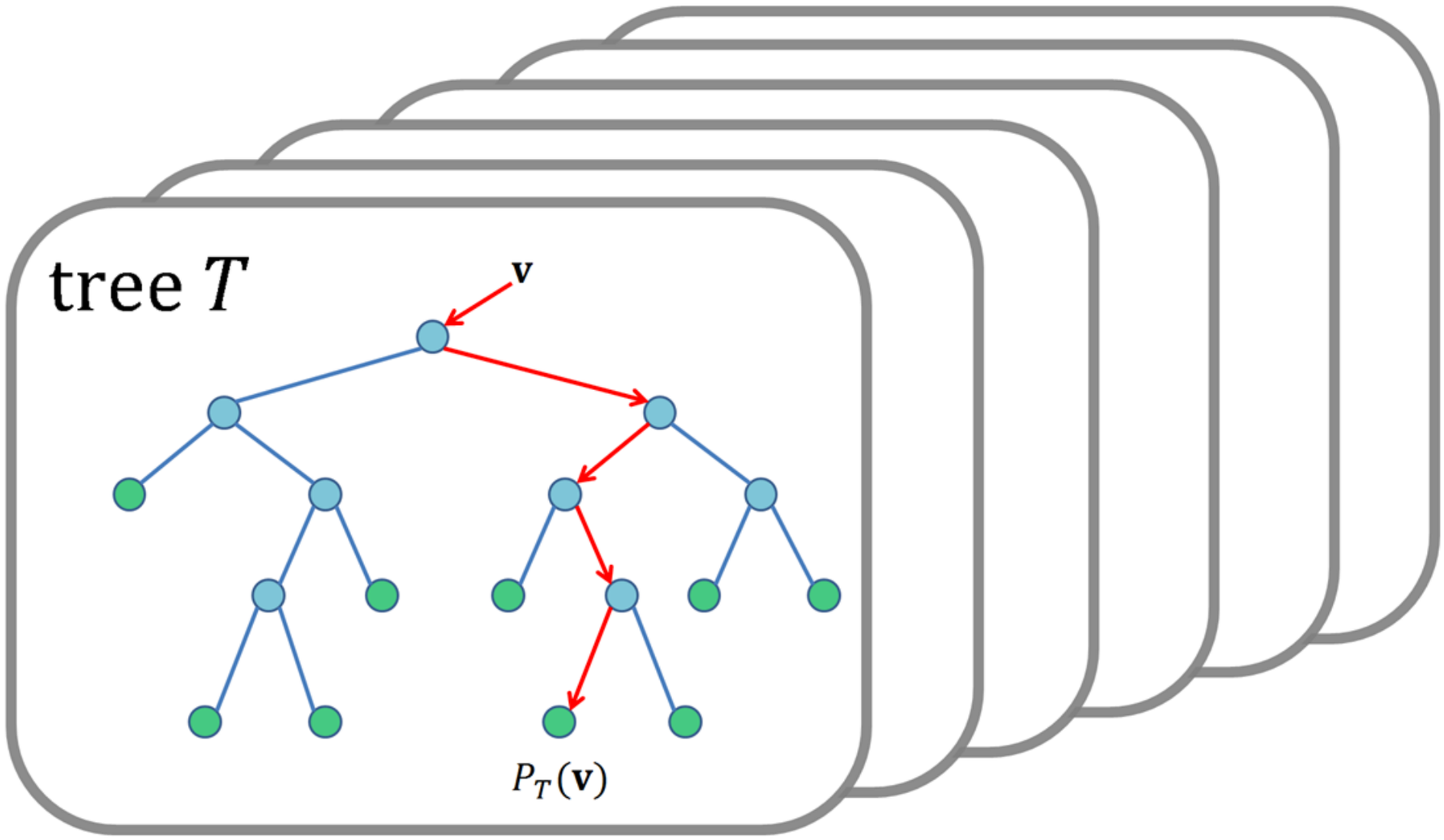
Clustering: to find subtypes among predicted outcomes

Predicted outcomes

k-means clustering



Random forest used to determine important predictors



Rationale of random forest

Ensemble of learning models corrects for propensity of decision trees to overfit.

While the predictions of a single tree are highly sensitive to noise in its training set, the average of many trees is not, as long as the trees are not correlated.

Bagging / bootstrap sampling de-correlates trees, i.e. decreases model variance without increasing bias.

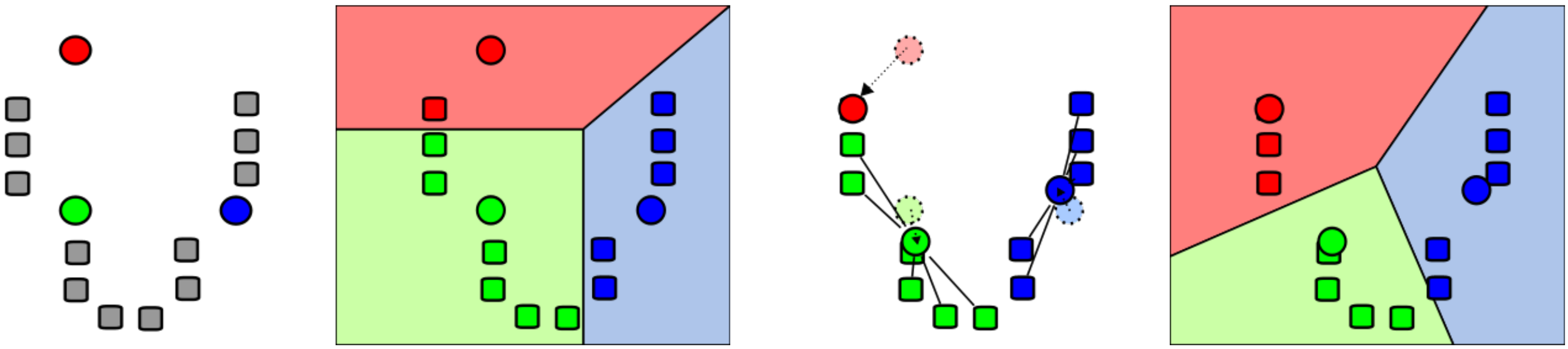
Lasso GLM removes redundant (i.e. correlated) predictors

$$\min_{\beta_0,\beta}\left(\frac{1}{N}\text{Deviance}\left(\beta_0,\beta\right)+\lambda\sum_{j=1}^{p}|\beta_j|\right)$$

Model fit

Penalty for correlated predictors

K-means clustering



Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Four outcomes of this study

1. # years since onset of episode lasting 2+ wks*: 13.0% [6.2 - 29.4%]
2. # years since onset of episode lasting *most days* of the year* 0.0% [0.0 - 9.3%]
3. Hospitalization from depression: 4.3%
4. Disability because of depression: 1.6%

*transformed to continuous data by dividing by # years between age at interview and AOO+1

Random forest (“recursive partitioning”) identified interactions for GLM

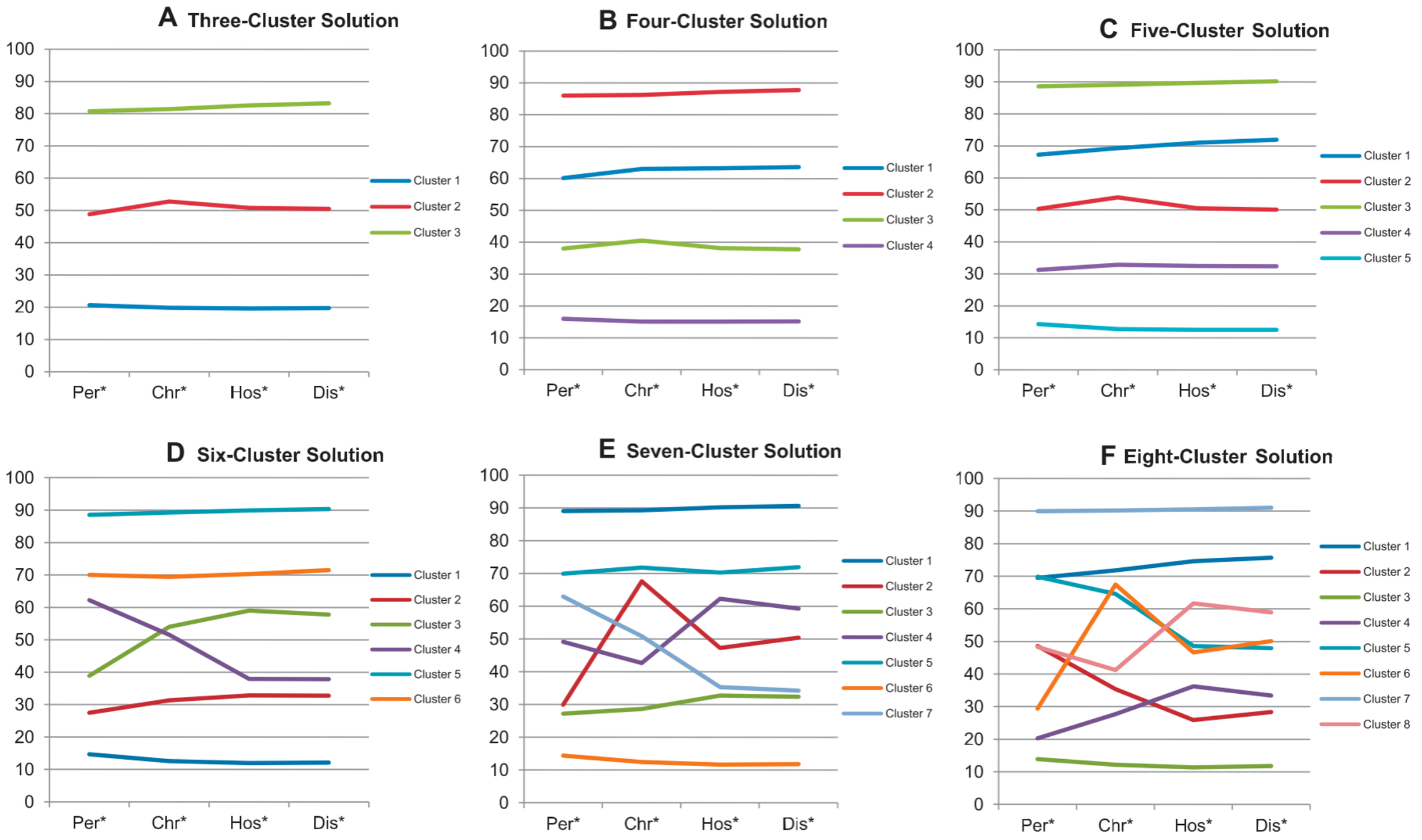
“the terminal nodes repeatedly predicting outcomes … all involved two-way or three-way interactions between child-adolescent AOO, suicidality, and anxiety during index depressive episodes.”

“all two- and three-way interactions among AOO, anxiety, and suicidality were included in the Lasso GLMs.”

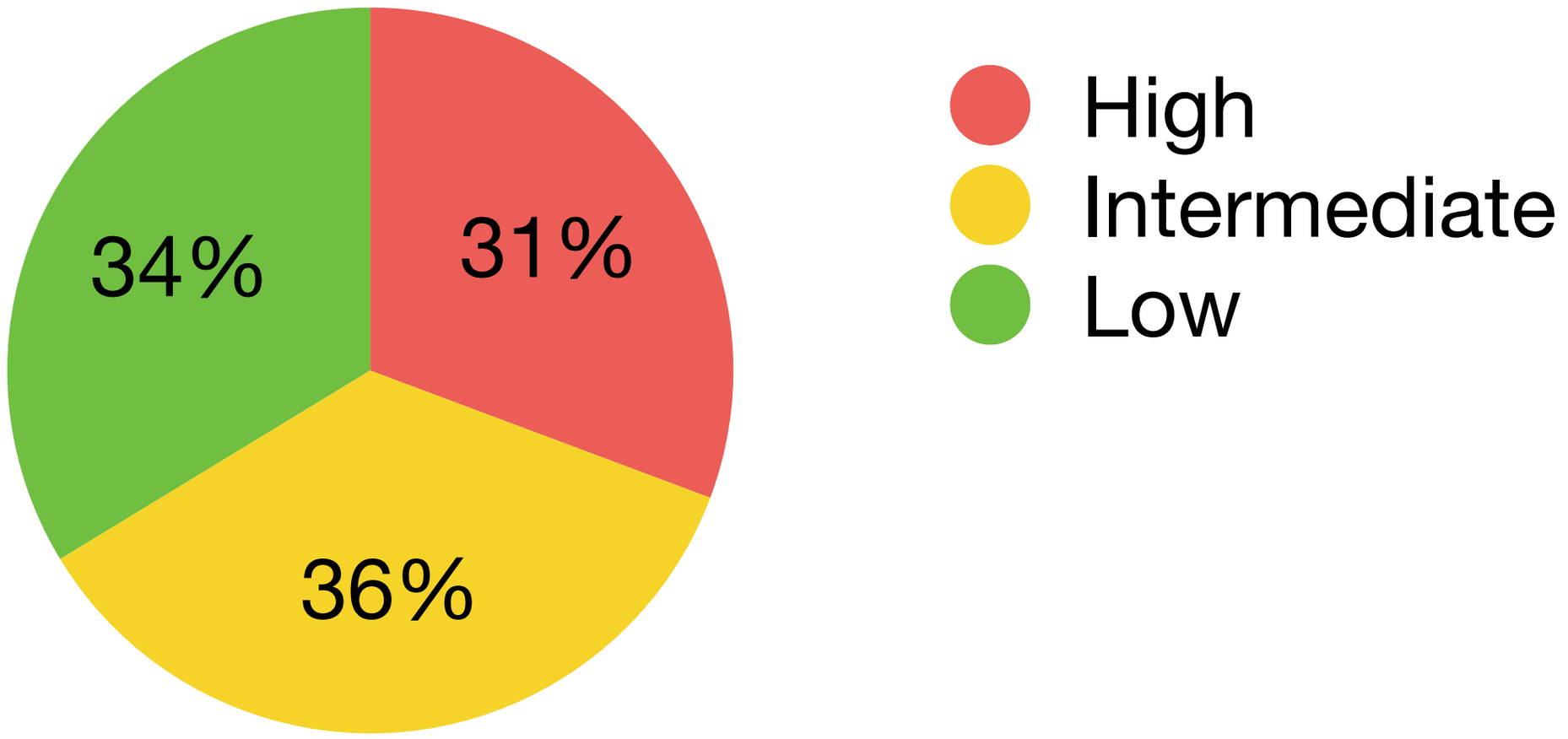
Lasso GLMs retained predictors

4: persistence
8: chronicity
11: hospitalization
11: disability

Clustering predicted outcome scores



Cluster analysis resulted in three outcomes risk groups:



Takeaways from table 3 (assocations of cluster membership w/ positive screening characteristics:

Tautology to show high-risk cluster has higher RR compared to other clusters

Very low PPV (% of respondents in high-risk cluster that experienced adverse outcome)

Moderate sensitivity (% of adverse outcomes that occurred in the high-risk cluster)

Overall results: 1/3

Recursive partitioning found an early-onset anxious-suicidal subtype associated with all four outcomes, and a late-onset anxious-suicidal subtype associated with chronicity.

Overall results: 2/3

GLMs found the # of index episode symptoms were significant predictors of all outcomes.

“The most consistent and powerful of these was severe dysphoria…”

Overall results: 3/3

Strong clustering was found in these predicted values across outcomes, with 30% of the high-risk cluster accounting for >67% of cases with multiple indicators of high long-term persistence, chronicity, and severity.

Overview

1. Quick summary of paper
2. Major depressive disorder
3. Data
4. Methods
5. Results
6. Critique

Critique

Retrospective data; structured interviews

No data on comorbidity or treatment status

Focus on just index episode symptoms

Recursive partitioning requires really big N

Rather poorly written…