# Predicting Strokes From Health Histories Using Relational Random Forests

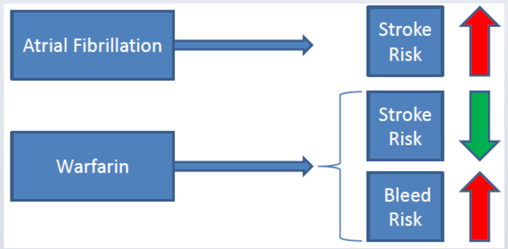
Zach Shahn (zss2101@columbia.edu)*, Patrick Ryan**, and David Madigan*

*Columbia University, **Observational Medical Outcomes Partnership

## The General Problem

**How do we identify informative complex temporal relations among multiple health events and incorporate them into predictions?**

### A Specific Example: Predicting Strokes in Atrial Fibrillation Patients




**Goal: Identify patients with sufficiently low stroke risk to safely be spared warfarin.**

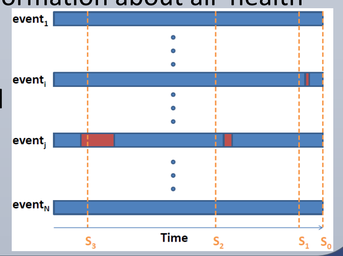
### CHADS2 (Sometimes Used Clinically)

| Predictor | Point Value |
|---|---|
| Congestive Heart Failure | 1 |
| Hypertension | 1 |
| Age > 75 | 1 |
| Diabetes Mellitus | 1 |
| Stroke | 2 |
| Score = CHF + Htn + Age + T2DM + Stroke | |

| Total Score | Estimated Probability of Stroke in Next Year |
|---|---|
| 0 | .019 |
| 1 | .028 |
| 2 | .04 |
| 3 | .059 |
| 4 | .085 |
| 5 | .125 |
| 6 | .182 |

### Standard Machine Learning Approach

We made ~100K binary predictors containing course temporal information about all health events. We fed these into large scale L1-regularized logistic regression and random forest algorithms.




## Relational Random Forests (RRF)

Idea: Generate informative temporal patterns involving multiple health events at the nodes of randomized decision trees. Classify stroke patients based on the presence or absence of these patterns, i.e. classify by terminal node membership as in a standard random forest.

### Node Splitting Algorithm

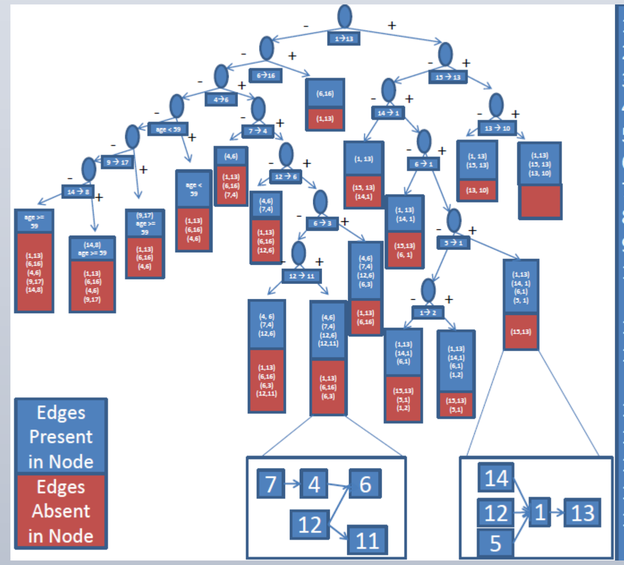
**Definition 1**: An 'edge' is a triplet **($e_1$,$e_2$,R)** where **$e_1$, $e_2$** are health events and **R** specifies a temporal relation that holds between them. (e.g. Asthma is diagnosed between 20 and 50 days before Diabetes)

**Definition 2:** A 'node' **N** in a tree is defined by two sets of edges, **I** and **E.** **N** consists of the patients in whom every edge in **I** and none of the edges in **E** are present .

To split node **N** defined by edge sets **I** and **E**:

- **M** times: Select a random edge **r** from the health history of a random patient in **N** such that **r** is connected to **I** and not contained in **E**.
- Split **N** on the most discriminating edge generated in the previous step.

### One Sample Tree (of Thousands)



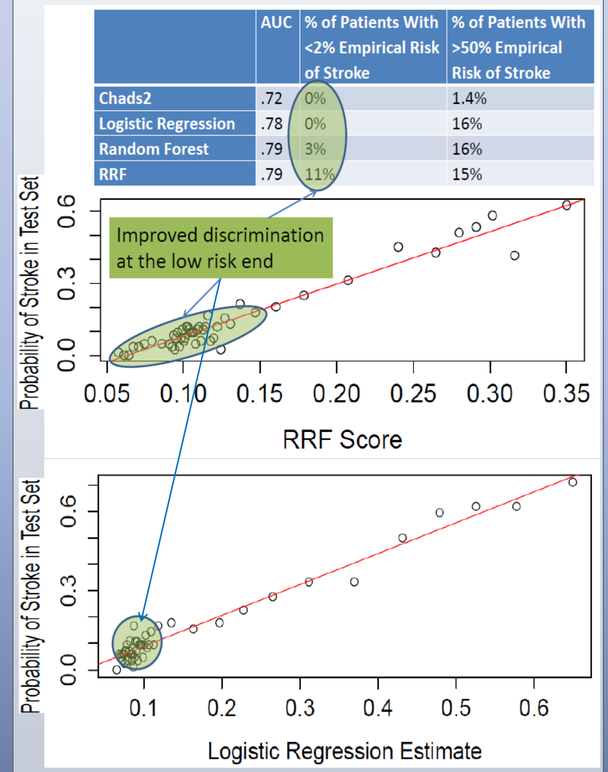

1.Vascular disorders NEC
2.Central nervous system vascular disorders NEC
3.Total fluid volume increased
4. Vascular disorders
5.Cardiac failure congestive
6.Nervous system disorders
7.Respiratory System
8.Eye disorders
9.Coronary artery disorders NEC
10. Haematological and lymphoid tissue therapeutic procedures
11.Anti-inflammatory & Antirheumatic
12.Non-steroids drugs for obstructive airway diseases
13.Blood and blood forming organs
14.Antithrombotic agents
15.Opioids
16.Myocardial disorders NEC
17.Arteriosclerosis, stenosis, vascular insufficiency and necrosis

The labeled connected graphs beneath the tree represent temporal patterns indicated by two terminal nodes. For example, the left graph conveys: event 7 occurs before event 4, event 4 occurs before event 6, event 12 occurs before event 6, event 12 occurs before event 11.

## Results

We used various methods to estimate probability of suffering a stroke within a year for patients with Afib in the OMOP Medicaid database.

| | AUC | % of Patients With <2% Empirical Risk of Stroke | % of Patients With >50% Empirical Risk of Stroke |
|---|---|---|---|
| Chads2 | .72 | 0% | 1.4% |
| Logistic Regression | .78 | 0% | 16% |
| Random Forest | .79 | 3% | 16% |
| RRF | .79 | 11% | 15% |




RRF was the only method able to discriminate between patients at the lower tail of stroke risk and identify a sizable population that can safely be spared warfarin.

### Conclusion

RRF can efficiently incorporate information from the vast space of temporal interactions among multiple variables into its predictions, making it a promising tool when covariates are a high dimensional time series.