Some Applications of Machine Learning algorithms in Pharma

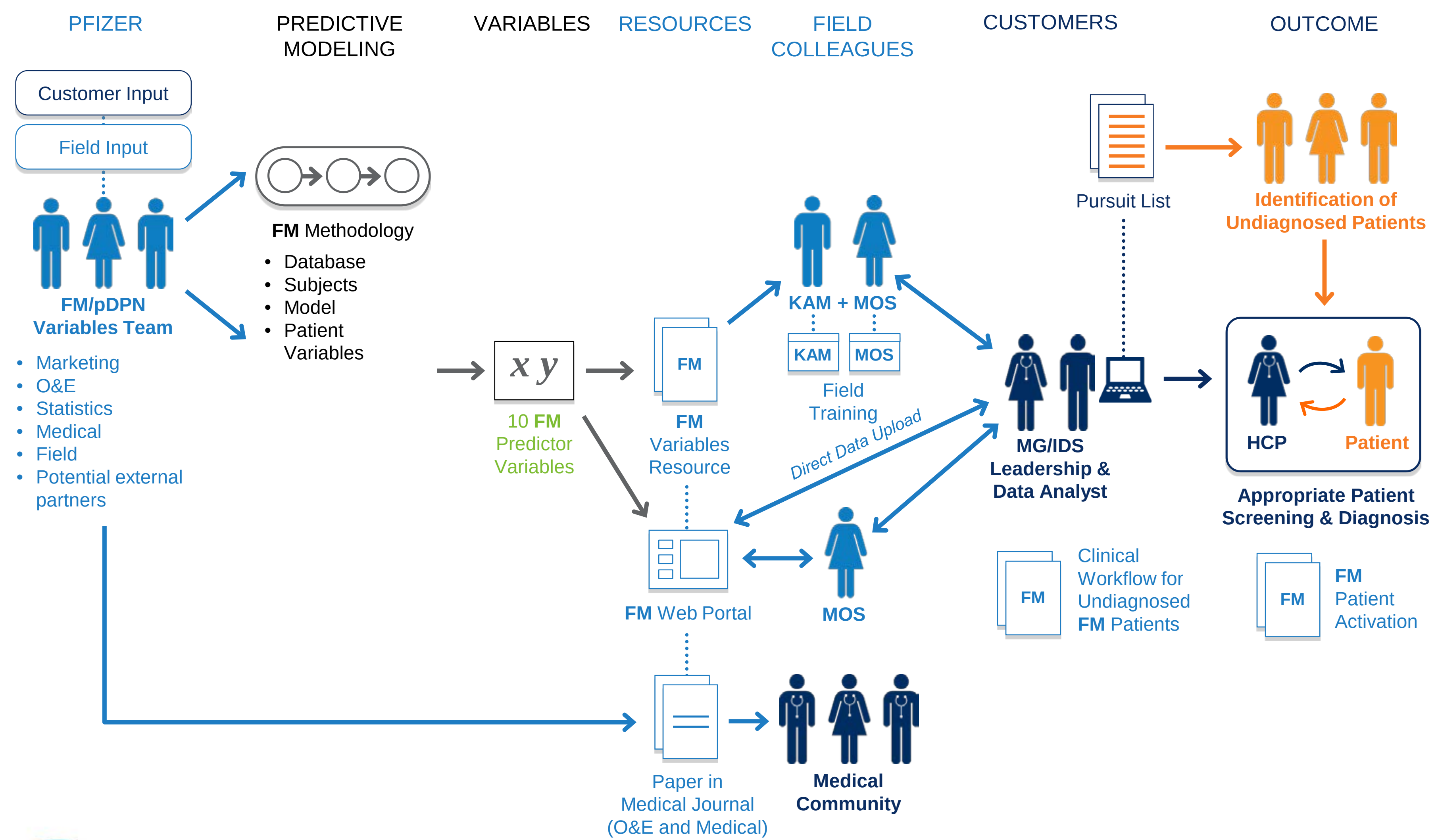
Birol Emir

April, 2018 – NISS-Merck Meet Up

Outline

- Early identification of Fibromyalgia (FM) patients using EM
  - with Jack Mardekian & Max Kuhn
- A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)
  - with Kjell Johnson & Max Kuhn
- A Wide and Deep Learning application to identify CV events from ER Claims data
  - With Pfizer and Optum Colleagues
- ?s

Overall Analytical Process



Topic

- **Early identification of Fibromyalgia (FM) patients using EM**
  - **with Jack Mardekian & Max Kuhn**
- A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)
  - with Kjell Johnson & Max Kuhn
- A Wide and Deep Learning application to identify CV events from ER Claims data
  - With Pfizer and Optum Colleagues
- ?s

Matthew Herper
Forbes Staff
FOLLOW

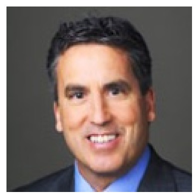
I cover science and medicine, and believe this is biology's century.

PHARMA & HEALTHCARE 2/17/2015 @ 10:18AM | 5,675 views

How Pfizer Is Using Big Data To Power Patient Care

Published in Forbes

+ Comment Now + Follow Comments



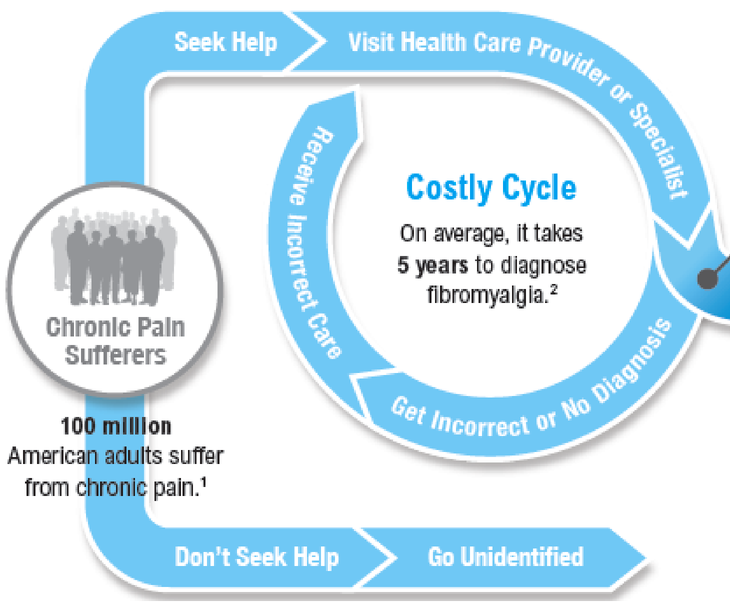
GUEST POST WRITTEN BY
Geno Germano
Group President, Global Innovative Pharma Business, Pfizer

"While working toward this vision of connected care, data are already informing ho conditions are diagnosed and managed today. For example, people with the chron condition called fibromyalgia which causes widespread pain, fatigue and cognitive issues, can cycle from doctor to doctor for up to five years before getting an accurate diagnosis. Using a large Electronic Medical Record database of de-identified patient data, and what we know about fibromyalgia from the medical literature, we've created a model to help clinicians identify patients that might be suffering from fibromyalgia earlier so patients can get effective care. [Editor's note: Germano's company, Pfizer, sells Lyrica, a drug to treat fibromyalgia pain.]"

Fibromyalgia in Context

Fibromyalgia Patient Journey

PRE-DIAGNOSIS



100 million American adults suffer from chronic pain.[1]

*Data from health care provider and consumer Pfizer market research.

DIAGNOSIS

How might we help **break the costly cycle** and ensure that potential fibromyalgia patients are identified, appropriately diagnosed, and managed?

Symptoms linked to fibromyalgia may include:

- Uncontrolled pain
- Co-morbid anxiety and/or depression
- Irritable bowel syndrome (IBS)
- Chronic fatigue

Over **5 million** people suffer from fibromyalgia,[3] but only **36%** of fibromyalgia patients are correctly diagnosed.[4*]

MANAGEMENT

Patient and health care provider work together to develop a treatment plan which may include:

- Patient education
- Setting treatment goals
- Applying a multimodal treatment approach
- Tracking progress

Model Parameters and Implementation

Study objective: Develop predictive models of fibromyalgia diagnosis to potentially facilitate earlier diagnosis and treatment through use of real-world data.

Database and Issues: Electronic Health Records (EHR) data from the Humedica database

- 587,961 patients meeting inclusion and exclusion criteria
- Train/Test ¾ vs ¼

	Parameters
Method	Random Forest with 1500 bootstraps
Class Imbalance	Internal down sampling
mtry	13
CV	10-fold repeated 5 times

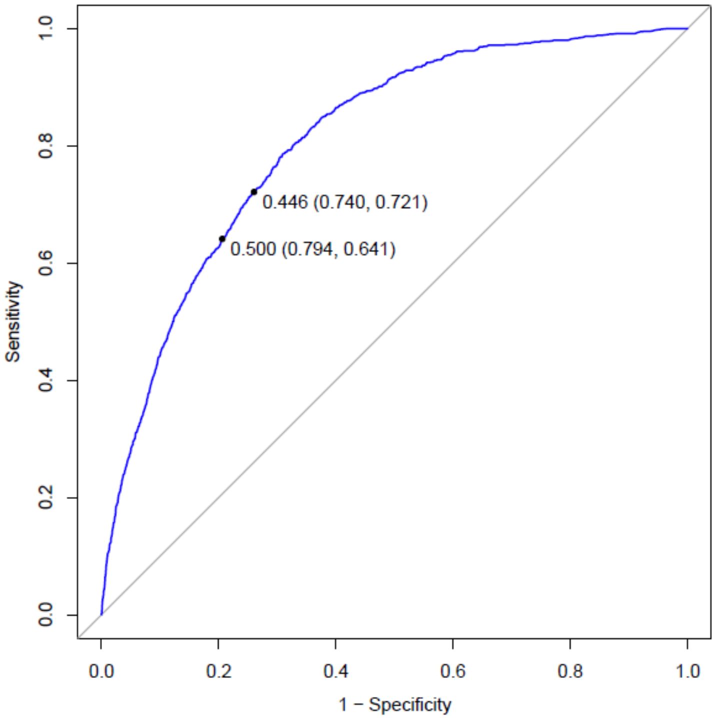
Training Data Results

The final analysis on the training data set incorporated the top 10 predictor variables that were suggested by the random forest model, ranked by their importance (normalized to 100%) based on the variable with the largest loss in prediction performance by its omission in the model.

	Cut	ROC	Sen	Spe
DS:	0.5	0.824	0.687	0.796
	0.446	0.824	0.757	0.741

Test Data Results

Data: evalResults$RF in 145930 controls) < 1055 cases
Area under the curve: 0.8097



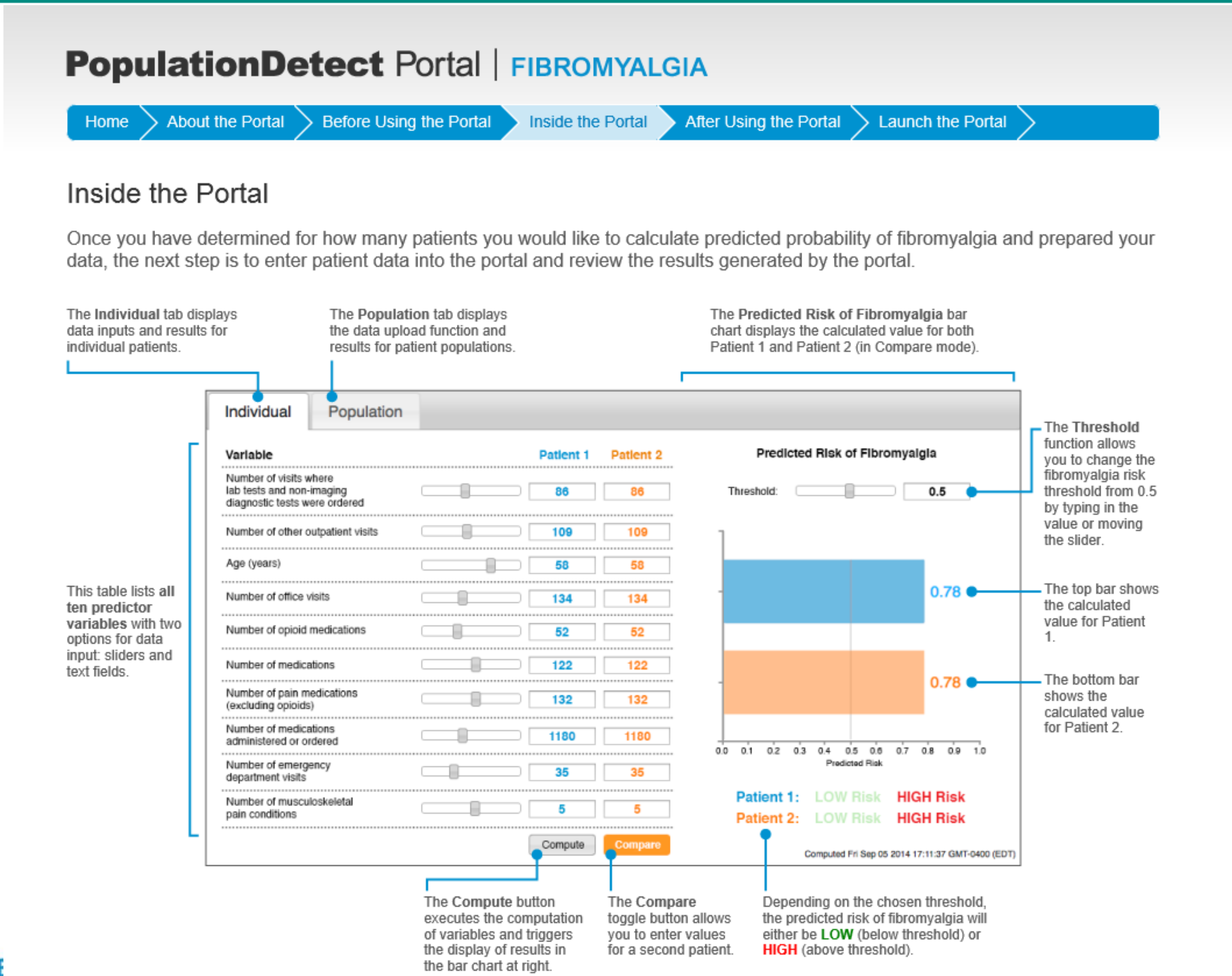

```
CUT OFF 0.5

Confusion Matrix and Statistics

          Reference
Prediction     FM   noFM
     FM       676  30044
     noFM     379 115886

               Accuracy : 0.793
                 95% CI : (0.7909, 0.7951)
    No Information Rate : 0.9928
    P-Value [Acc > NIR] : 1

                  Kappa : NA
 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.640758
            Specificity : 0.794120
         Pos Pred Value : 0.022005
         Neg Pred Value : 0.996740
             Prevalence : 0.007178
         Detection Rate : 0.004599
   Detection Prevalence : 0.209001
      Balanced Accuracy : 0.717439

       'Positive' Class : FM
```

Communication and use of this model

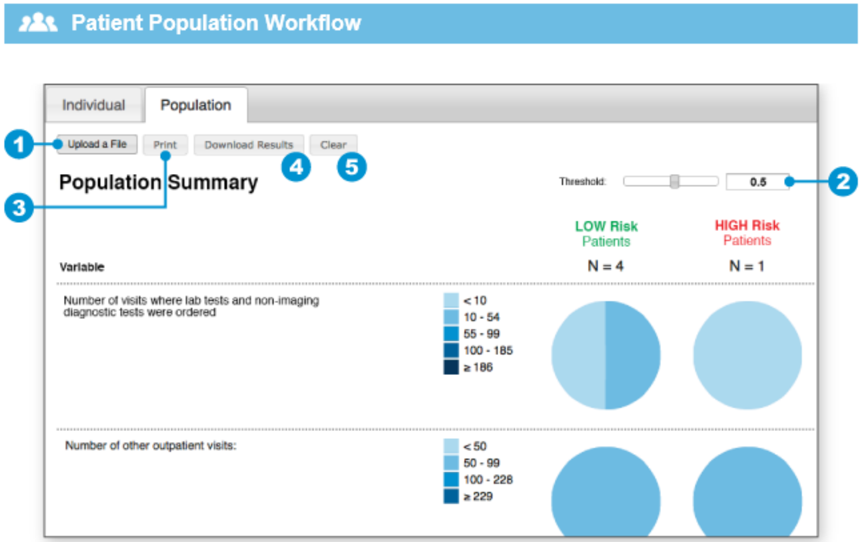
- RF is a complex model with a large footprint
- It is not interpretable
- For future prediction you need the full trees
- How can we make this a practicable for prediction purposes

WEB PORTAL with a nice interface

Outsourced Portal



Population Workflow



When looking at a **population** of patients…

1. Click on the "Upload a File" button to select the CSV file you have prepared for upload. Results will be generated if the file upload is successful.
2. **Optional:** Use the "Threshold" slider or text field to adjust the threshold between "LOW Risk" and "HIGH Risk."
3. Click on the "Print" button to print the page.
4. Click on the "Download Results" button to download a CSV file containing the predicted probabilities of fibromyalgia for the population.
5. To clear the results and start over, click on the "Clear" button.

Rules Using C5.0 for the manuscript

- To generate these rules, a simulated dataset was created in order to obtain a broader range of values for the ten predictors and to avoid concerns of overfitting through repeated use of the training dataset. The minimum, maximum, 20th, 40th, 60th, and 80th percentiles of the ten predictors identified by the random forest model were computed using the training dataset.
- The simulated dataset was run through the random forest model to obtain a predicted probability of an FM diagnosis for each patient.
- Focusing on the simulated patients with the highest (> 0.70) and lowest (< 0.20) predicted probabilities of FM resulted in 4,179 simulated patients for analysis and the C5.0 rules were then applied to classify these patients.

Rules

Table 3 Rules for identifying FM and no-FM subjects based on results of the predictive modeling using a technique known as C5.0 rules

Rule number	Predictive class	Rule (all components must be met)	Number of subjects predicted in simulated dataset (n=4,179) to belong to predictive class	Percentage of subjects in simulated dataset (n=4179) correctly identified in predictive class	Sensitivity (%) computed in patients identified by rule applied to test dataset (n=146,985)	Specificity (%) computed in patients identified by rule applied to test dataset (n=146,985)
1	FM	Number of outpatient visits >0 Number of prescriptions administered ≤3 Number of musculoskeletal pain conditions >0	308	99.7	78.3	39.7
2	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of musculoskeletal pain conditions >0	247	99.6	85.6	26.6
3	FM	Number of outpatient visits >0 Number of office visits ≤9 Number of opioid prescriptions >0	208	99.5	75.9	34.9
4	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of emergency room visits >0	102	99	94.8	15.4
5	FM	Number of visits where laboratory/non-imaging tests were ordered >0 Number of pain medications excluding opioids >2	63	98.5	92.7	18.5
6	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of opioid prescriptions =0 Number of musculoskeletal pain conditions =0	2,176	100.0	99.6	0
7	No-FM	Number of opioid prescriptions =0 Number of pain medications excluding opioids ≤2 Number of emergency room visits =0 Number of musculoskeletal pain conditions =0	1,761	99.9	96.6	5.6
8	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of office visits >9	1,224	99.9	94.7	36.3
9	No-FM	Number of visits where laboratory/non-imaging tests were ordered =0 Number of outpatient visits =0	3,091	99	98.2	15.8

Topic

- Early identification of Fibromyalgia (FM) patients using EM
  - with Jack Mardekian & Max Kuhn

- **A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)**
  - **with Kjell Johnson & Max Kuhn**

- A Wide and Deep Learning application to identify CV events from ER Claims data
  - With Pfizer and Optum Colleagues

- ?s

Model Parameters and Implementation

Study objective: This post hoc analysis used 8 predictive models to evaluate potential predictors of achieving at least 50% pain reduction by week 6 after treatment initiation with pregabalin

Database and Issues: This study was a 6-week, prospective, non-interventional, drug-monitoring study of patients who Were treated with pregabalin for NeP from 2004 through 2005 in Germany

- 15,301 patients
- To adjust for the high imbalance in the responder distribution (75% of patients were 50% responders)
- Train/Test Split - 1000 training 1000 test

Method
LDA
RPart
CTree
k-NN
RF
GBM
SVM
NB

	Parameters
CV	10-fold repeated 5 times

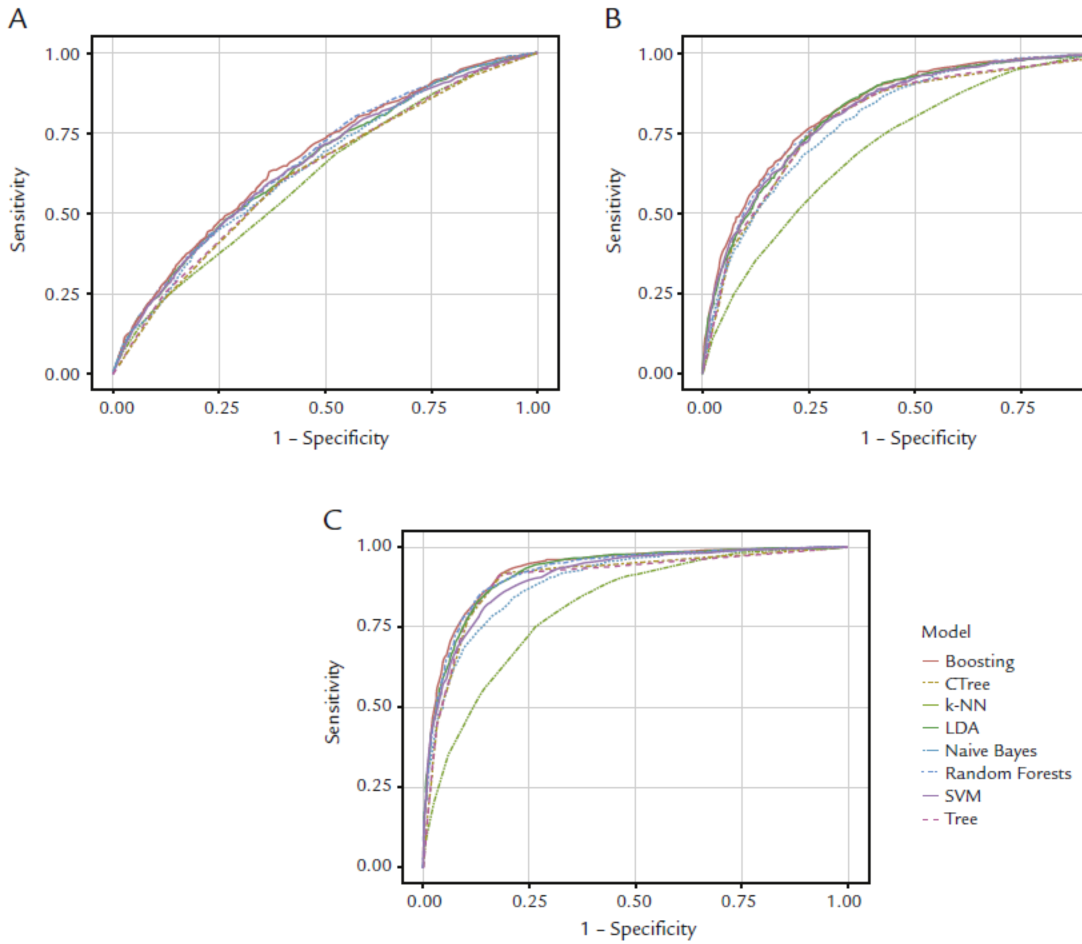
A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)

- Baseline demographic and clinical characteristics were evaluated for 46 potential predictors. Post baseline pain information at treatment weeks 1 and 3 was also available.

Table IV. Variable importance for the internal balanced training set including baseline predictors and pain response at weeks 1 and 3.

Potential Predictor	ROC*	RPart*	PLS*	RF*	GBM*	Average Importance
Pain change from week 3	100.0	100.0	100.0	100.0	100.0	100.0
Pain change from week 1	76.1	53.4	71.9	36.1	3.1	48.1
Baseline NRS pain score	0.7	24.0	7.2	19.1	19.7	14.1
Depression	19.3	15.6	27.7	4.9	1.9	13.9
Pregabalin as monotherapy	21.4	7.3	25.4	3.9	0.6	11.7

A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)



Method	Accuracy (95% CI)
LDA	0.89 (0.88–0.90)
RPart	0.84 (0.83–0.85)
CTree	0.83 (0.82–0.85)
k-NN	0.84 (0.83–0.85)
RF	0.88 (0.87–0.89)
GBM	0.88 (0.87–0.89)
SVM	0.86 (0.85–0.87)
NB	0.84 (0.82–0.85)

Figure. Receiver-operating characteristic curves for models, including: (A) baseline predictors, (B) baseline predictors and pain change from baseline at week 1, and (C) baseline predictors and pain change from baseline at weeks 1 and 3. CTree = conditional inference tree; k-NN = k-nearest neighbors; LDA = linear discriminant analysis; SVM = support vector machines.

Topic

- **Early identification of Fibromyalgia (FM) patients using EM**
  - **with Jack Mardekian & Max Kuhn**
- A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)
  - with Kjell Johnson & Max Kuhn
- **A Wide and Deep Learning application to identify CV events from ER Claims data**
  - **With Pfizer and Optum Colleagues**
- ?s

A Wide and Deep Learning application to identify CV events from ER Claims data

- Using deep learning, recommend a classification model on
  - Predicting who will have a cardiology-related emergency department utilization (cases) or not (controls)
  - Examining both prevalence and label noise impacts on model performance on curated datasets

- Optum EHR database including
  - claims,
  - clinical and
  - semi-structured data extracted from the unstructured clinical notes in EMR records using NLP.

A Wide and Deep Learning application to identify CV events from ER Claims data

Baseline datasets identified…
- to enable independent testing of impacts of each of these challenges on model performance

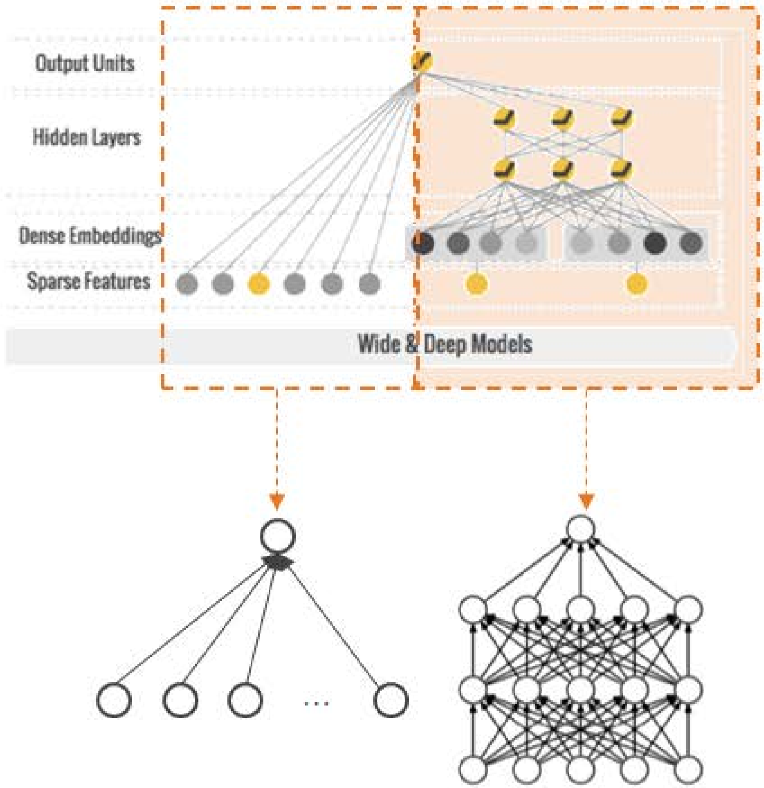
Dataset	Total Patients	Controls	Cases	Prevalence	Label Noise
ED Visit[2]	720,621	352,984	367,637	51%	0%
ED Visit	392,204	352,984	39,220	10%	0%
ED Visit	392,204	352,984	39,220	10%	10%
ED Visit	392,204	352,984	39,220	10%	20%
ED Visit	392,204	352,984	39,220	10%	30%

Volume – leveraging differently sized cohorts, further tested as an outcome of under-sampling for prevalence

Prevalence – under-sampling cases (true labels) to reduce prevalence (10%)

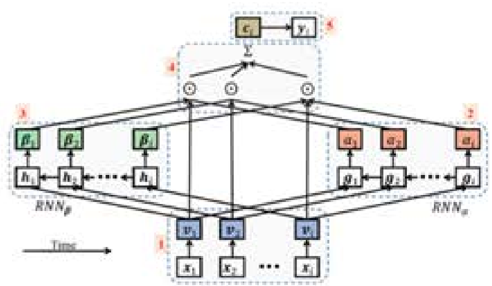
Label noise – selectively flipping labels (presence of ICD codes) to introduce noise (0, 10, 20, 30%)

ER claims data to predict CV events



FTRL[2]

FF



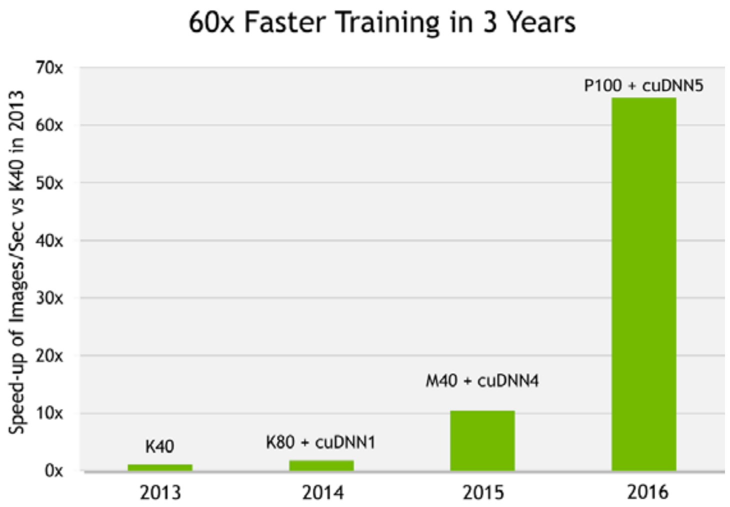
RETAIN[1]

1. RETAIN: An Interpretable Predictive Model for Healthcare using Reverse Time Attention Mechanism. E Choi, MT Bahadori, J Sun, J Kulas, A Schuetz, W Stewart Advances in Neural Information Processing Systems, 3504-3512
2. Follow the Regularized Leader: McMahan, H. Brendan et al. "Ad click prediction: a view from the trenches." *KDD*(2013).

Model Parameters and Implementation

Model evaluation is performed by splitting each dataset into three independent datasets:

- Training set – dataset comprised of a random selection of 60% of each dataset
- Validation set – held-out dataset comprised of a random selection of 20% of each dataset
- Test set – held-out dataset comprised of a random selection of 20% of each dataset respect to the validation set).



	Tuning Parameters
Number of hidden layers	2
Neurons or cells in each hidden layer	150 in first, 50 in second
Learning rate	lr=0.00001 with val_loss reduceOnPlateau by a factor of 0.2
Drop-out rate at each layer	0.4 after first hidden, 0.2 after second hidden layer
Early Stopping	Early Stopping val_loss min_delta=0.0001, patience=4

We are experimenting with a number of different data models for different structured data types for use with some or all of the three deep learning models, specifically:

- One-hot encoding of raw data – a transformation of categorical data into as many binary variables as there are categories (e.g., from color = red, green, blue to red = true, false green = true, false blue = true, false)
- Binning – stratifying continuous variables to reduce dimensionality and aggregate some values

Comparison of model performance vs. dataset

Test Results Deep Learning Model on Optum ER Data

Prevalence	Label Noise	Accuracy[1]	Precision[1]	Recall[1]	F1 Score[1]	AU-ROC
51%	~0%	0.7227	0.7688	0.6526	0.7060	0.7942
50% Cases oversampled from 10%	~0%	0.7237	0.7682	0.6566	0.7080	0.7965
50% Cases oversampled from 10%	~10%	0.7224	0.7678	0.6535	0.7061	0.7962
50% Cases oversampled from 10%	~20%	0.7215	0.7673	0.6518	0.7049	0.7951
50% Cases oversampled from 10%	~30%	0.7216	0.7653	0.6553	0.7060	0.7942

1. Performance results shown for a classification threshold of 0.5
2. No clinical validation of the labels (ICD codes for heart failure) performed on this data to determine quality of diagnosis code documentation

References

- Emir B, Masters ET, Mardekian J, Clair A, Kuhn M, Silverman SL. (2015). Identification of a potential fibromyalgia diagnosis using random forest modeling applied to electronic medical records. J Pain Res. 2015 Jun 10;8:277-88. doi: 10.2147/JPR.S8256. eCollection 2015. PubMed PMID: 26089700; PubMed Central PMCID: PMC4467741.

- Emir B, Johnson K, Kuhn M, Parsons B. (2017). Predictive Modeling of Response to Pregabalin for the Treatment of Neuropathic Pain Using 6-Week Observational Data: A Spectrum of Modern Analytics Applications. Clin Ther. 2017 Jan;39(1):98-106.

- Kuhn and Johnson (2012) Applied Predictive Modeling

Last Slide

BACK UP

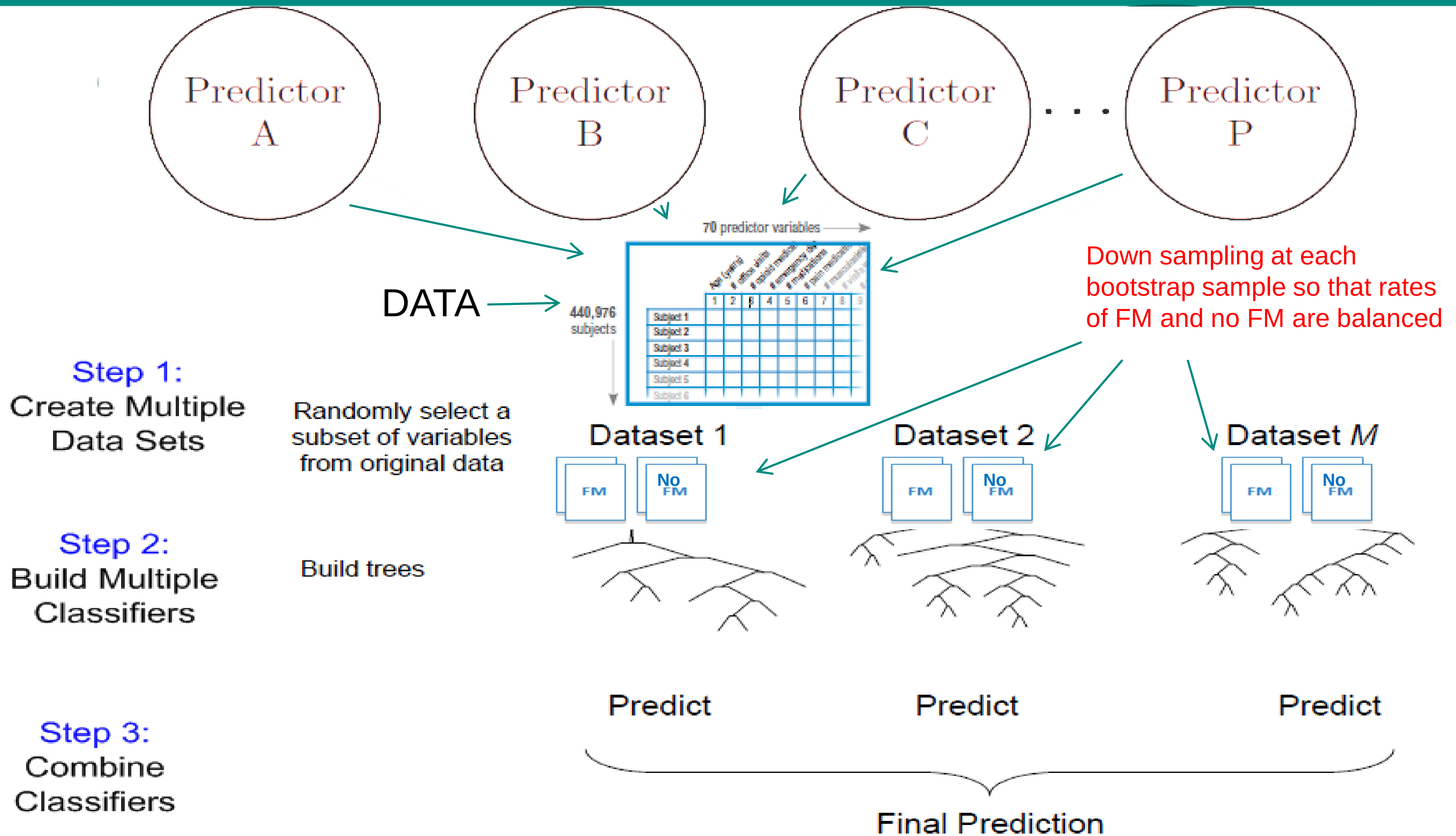
Transparent Reporting of a multivariable prediction model for individual Prognosis Or Diagnosis (TRIPOD)

www.tripod-statement.org

Tripod Checklist

Microsoft Word Document

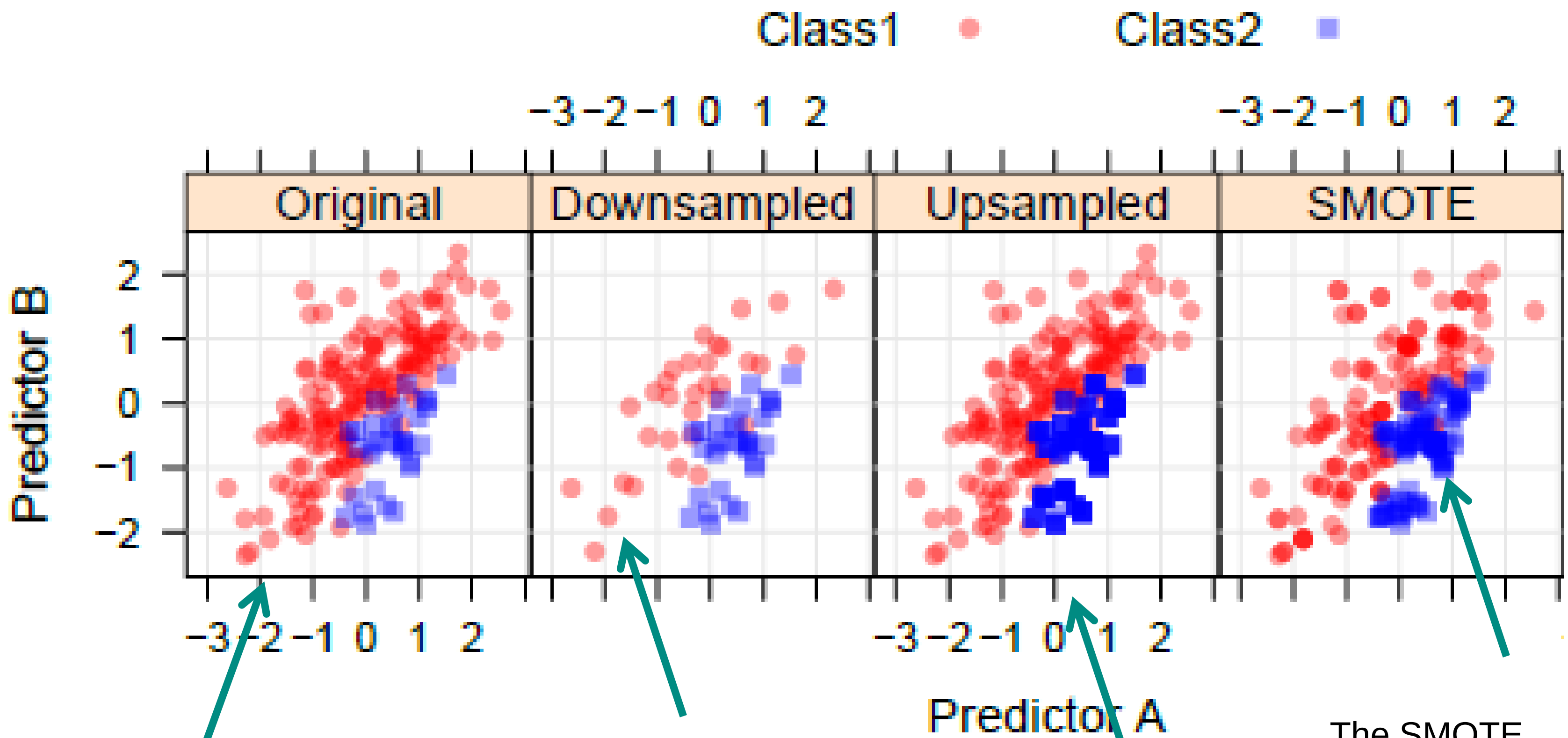
Random Forest



Breiman L. Random forests. Machine learning. 2001;45(1): 5-32.)

Majority wins from this ensemble

Some Remedies for Class Imbalance: Up Sampling, Down Sampling, & SMOTE



The original data contain 168 samples from the first class and 32 from the second class

The down-sampled version of the data reduced the total sample size to 64 cases evenly split between the classes.

The up-sampled data have 336 cases, now with 168 events.

The SMOTE version of the data has a smaller imbalance (with a 1.3:1 ratio) resulting from having 128 samples from the first class and 96 from the second

Kuhn and Johnson (2012)