Construction and Implementation of a Predictive Model to Identify Potential Fibromyalgia Patients

Birol Emir
Jack Mardekian

phuse SHARE
CONNECT SHARE ADVANCE

Artificial Intelligence and the Digital Transformation of Healthcare

July 18, 2019

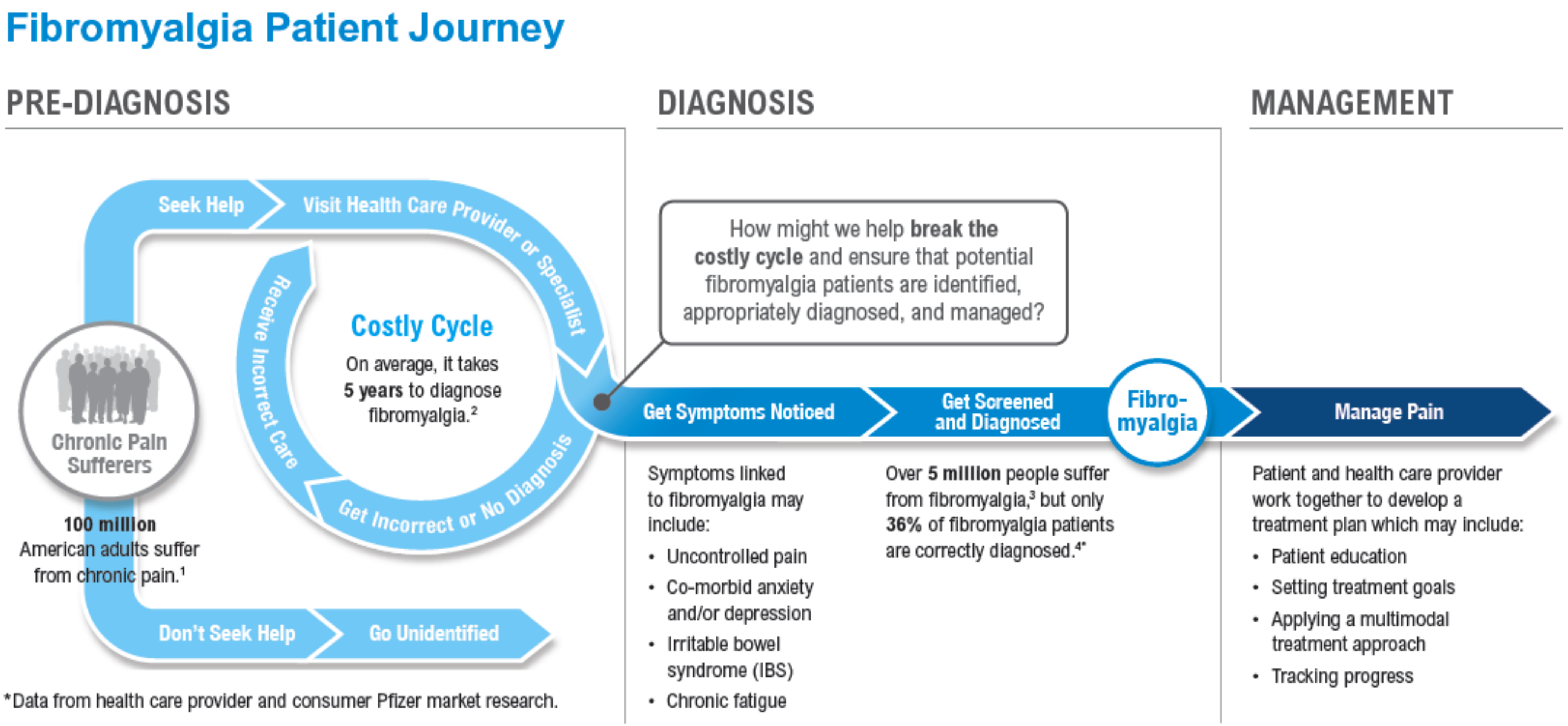
Pfizer BIOMETRICS & DATA MANAGEMENT
Global Product Development

TIME IS LIFE

Presentation Outline

- Project background
- Data Source
- Down Sampling
- Random Forest
- Communication and Implementation of Model
- Project update
- Questions

Fibromyalgia in Context



Study Objectives

To address the challenges associated with diagnosing fibromyalgia, Pfizer conducted a retrospective analysis to determine which patient variables could be associated with a fibromyalgia diagnosis.

The objective of this study is to utilize Electronic Health Records (EHR) data from the Humedica database to evaluate the characteristics associated with FM diagnosis. The objectives of the study include:

- **Summarize and describe subjects who develop fibromyalgia** and compare them to subjects who do not develop fibromyalgia in the EHR database.
- **Develop predictive models of fibromyalgia diagnosis** to potentially facilitate earlier diagnosis and treatment through use of real-world data.

Humedica Electronic Medical Records Database

- Has broad geographic representation across the United States
- Aggregates de-identified EMR data from healthcare providers across the continuum of care including hospitals, medical groups, and integrated delivery networks.
- Patient records are linked using a unique patient identifier and are fully compliant with the Health Insurance Portability and Accountability Act with regard to identification of patients and providers, as well as protected health information.

Humedica Key Clinical Content Areas

Sourced from Structured EMR Inputs

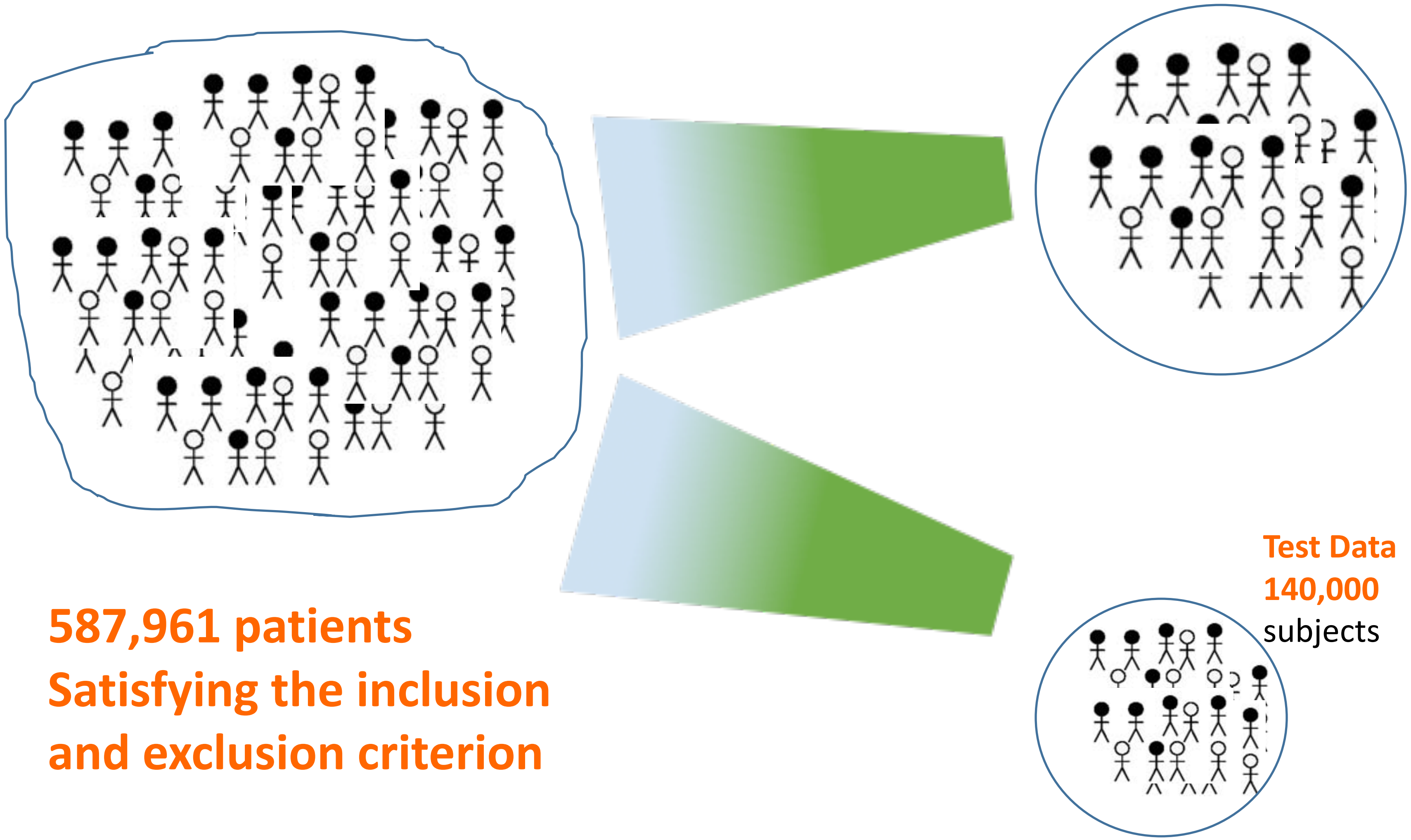
- Patient Information
- Provider Information
- Insurance Type
- Encounters
- Care Areas
- Visits (Summary Table of Inpatient, Observation, Emergency Patient)
- Medications – Written, Administered, and Patient-Reported
- Procedures
- Diagnoses
- Immunizations
- Microbiology Results
- Lab Results (200+ mapped labs)
- Clinical Observations (eg. Vital Signs, Body Measurements, Smoking Status)*

** Select Observations, like BMI, include additional information from provider notes that is tested against structured data and is validated on an on-going basis*

Sourced from Provider Notes

- Signs, Diseases, Symptoms (SDS)
- Drug Rationale
- Measurements
- Biomarkers

Optum (Humedica) EHR Data

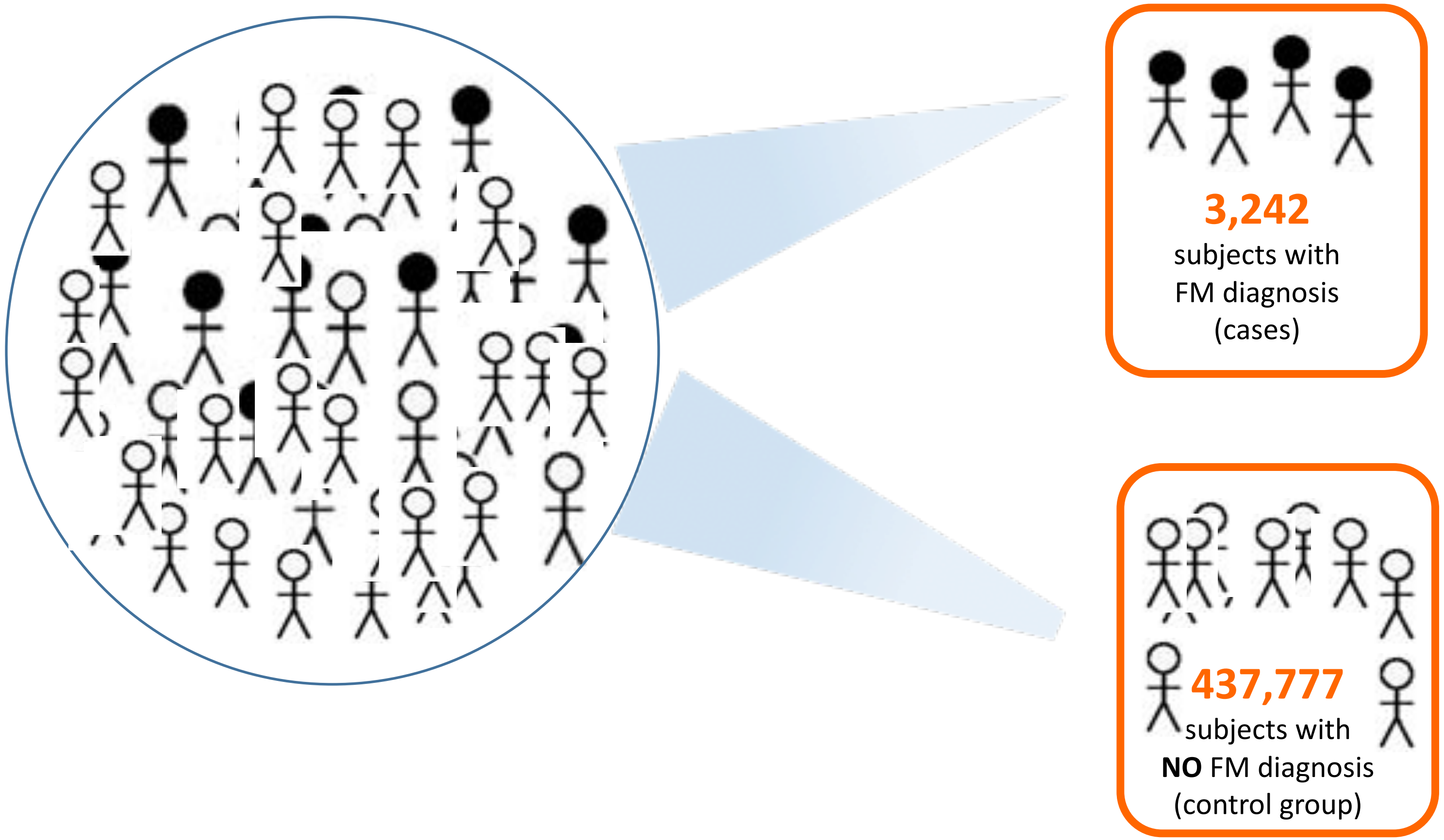


Training Data
441,019
subjects

Test Data
140,000
subjects

587,961 patients
Satisfying the inclusion
and exclusion criterion

Outcome Imbalance (Similar in Train and Test)



0.7%
FM

99.3%
no FM

Severe Class Imbalance

- An imbalance occurs when one or more classes have very low proportions in the training data as compared to the other classes.
  - In our data < 1% had FM diagnosis
- When modeling binary classes, the relative frequencies of the classes can have a significant impact on the effectiveness of the model; we should expect to see some adverse affects on the results as the model tries to deal both with the class imbalances and prediction simultaneously

Severe Class Imbalance (Continued)

- One consequence of this is that the performance is generally very biased against the class with the smallest frequencies.
  - For example, if the data have a majority of samples belonging to the first class and very few in the second class, most predictive models will maximize accuracy by predicting everything to be the first class. As a result there's usually great sensitivity but poor specificity.

Kuhn and Johnson (2012)

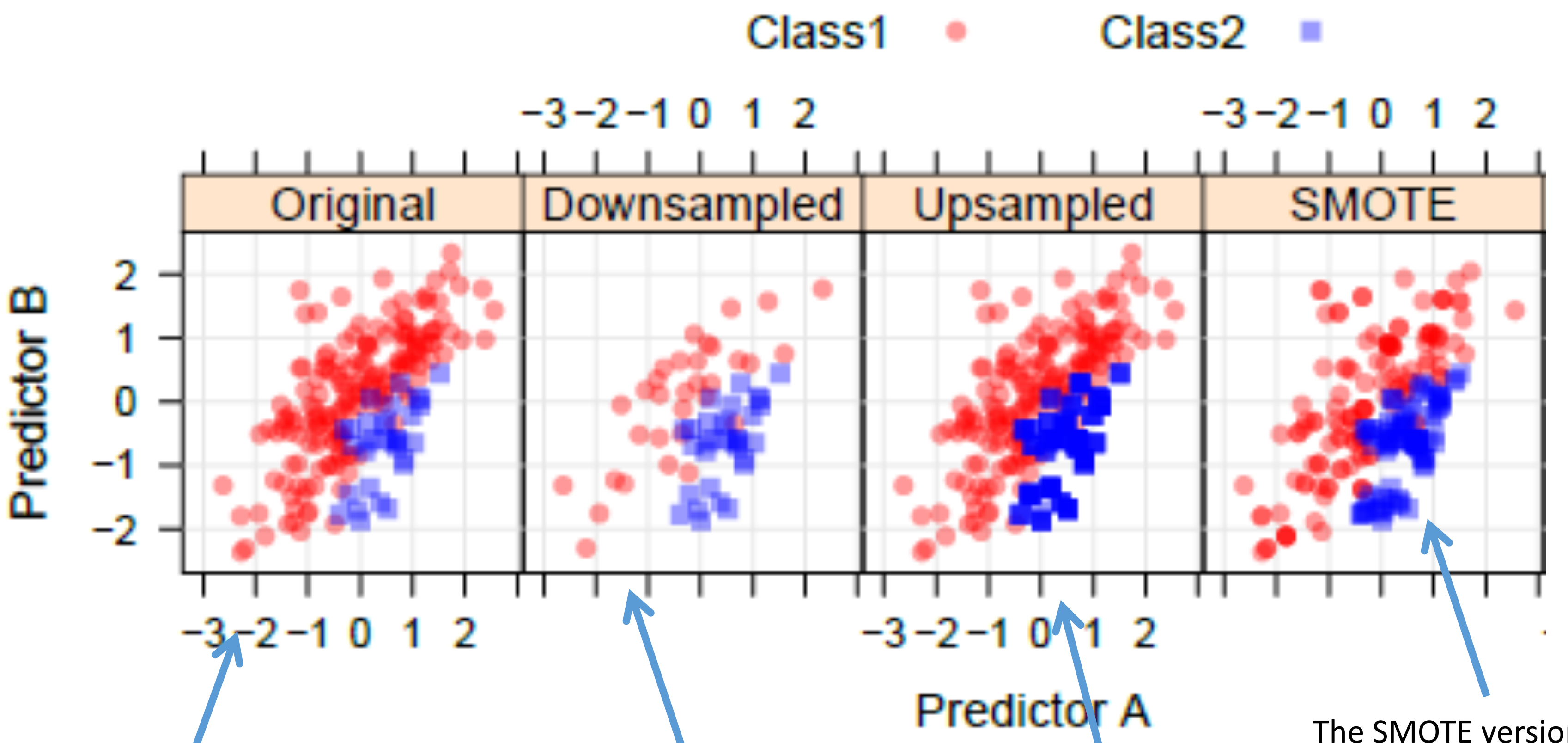
A Possible Solution

Two general *post hoc* approaches:

- Up-sampling: Any technique that simulates or imputes additional data points to improve balance.
- Down-sampling: Any technique that reduces the number of samples to improve the balance across classes.

Kuhn and Johnson (2012)

Down Sampling and SMOTE



The original data contain 168 samples from the first class and 32 from the second (a 5.2:1 ratio).

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)

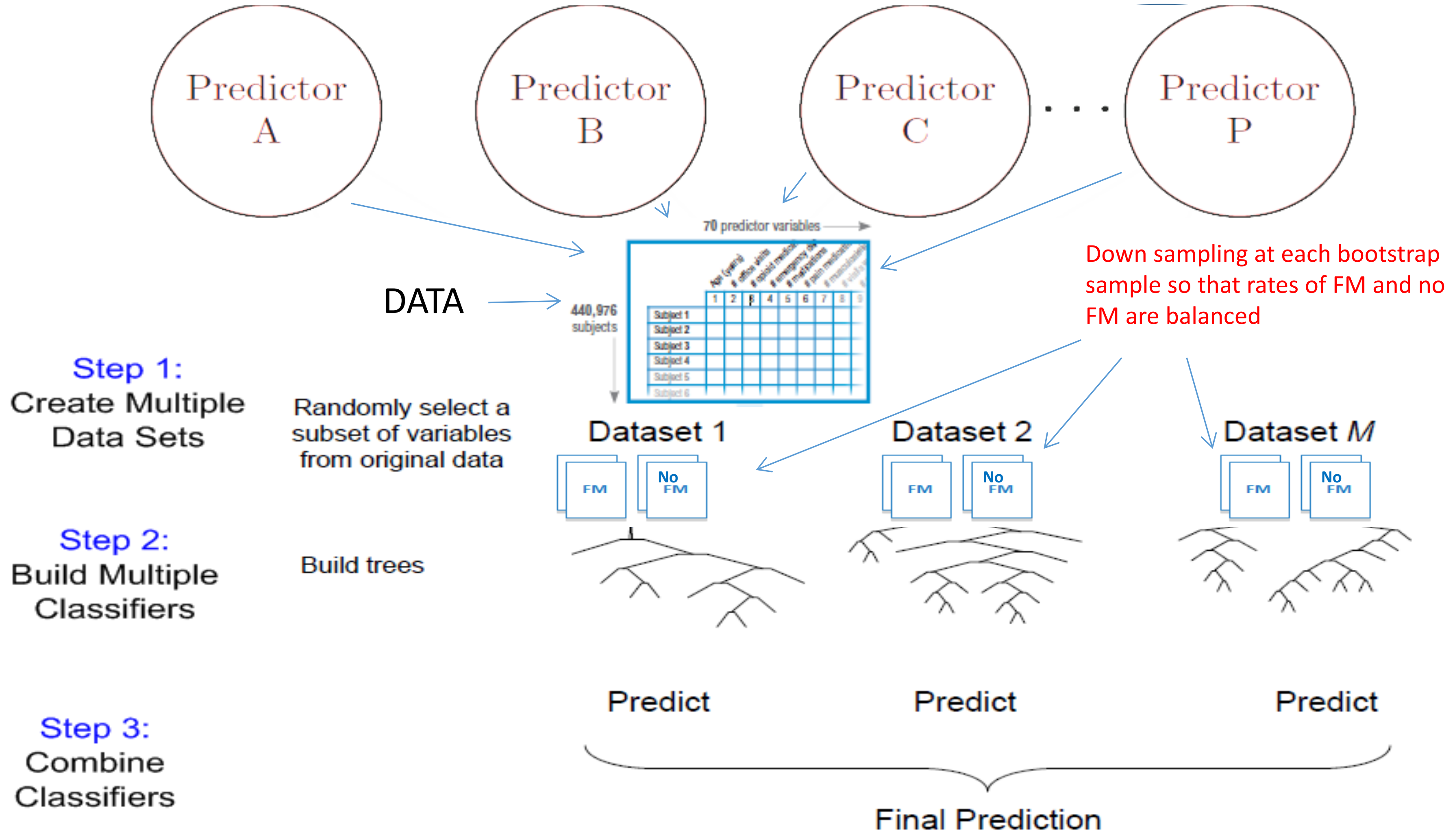
Random Forest (RF)

- Random forest modeling is a computationally extensive data mining technique that can be used to identify and rank the importance of predictors from among the range of input variables.
- This technique uses historical data from subjects and attempts to accurately predict future outcomes from classification trees generated through data resampling to produce an integrated prediction that can be highly accurate.

Random Forest on the Training Data

- RF can internally handle down sampling
- We fit a RF with 1500 bootstrap trees (full length), 10-fold cross validation repeated 5 times, with a strata variable so that a down sampling was done at each bootstrap, allowing the variable importance to be collected.
- The number of randomly selected variables was optimized based on the CV results. This whole process was repeated for different set of cutoff thresholds (other than the universal cutoff 50%).

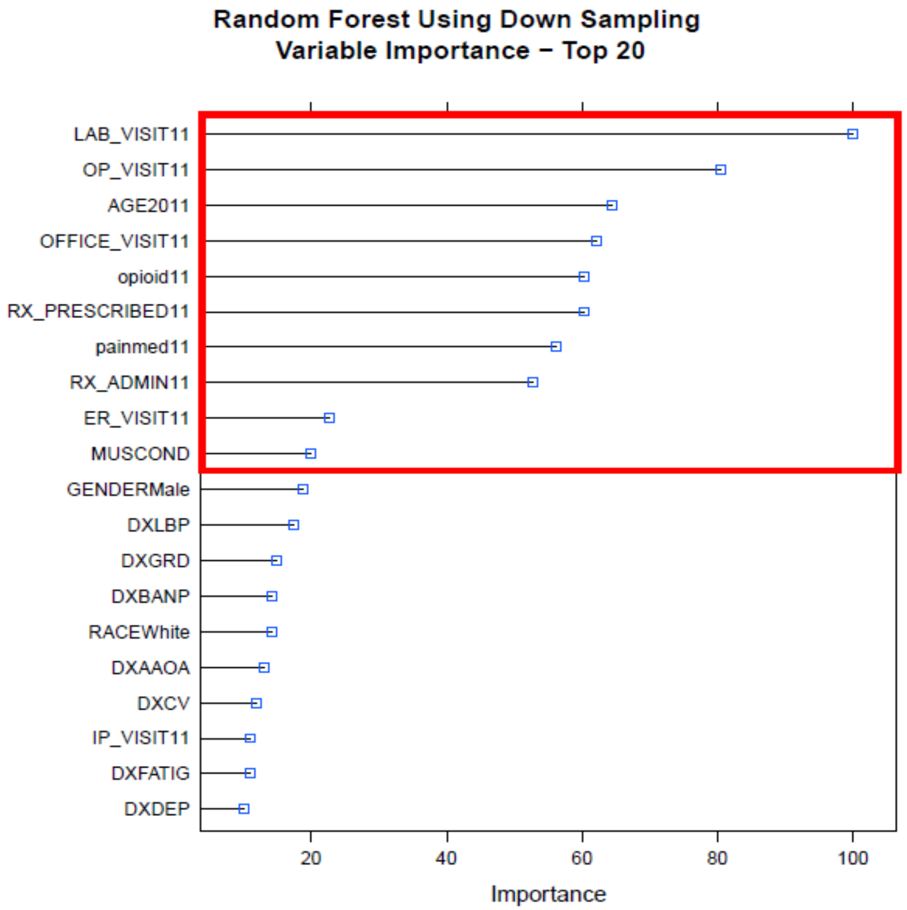
Overview of Random Forest



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

Majority wins from this ensemble

Variable Importance Random Forest



The variable importance plot shows the relative importance of each variable in the model. Variables were ordered by the loss of performance resulting from omission of the variable in the prediction model. The variable with the largest loss in predicting performance by its omission in the model is normalized to 100%. The plot above shows the top twenty variables with the top ten highlighted.

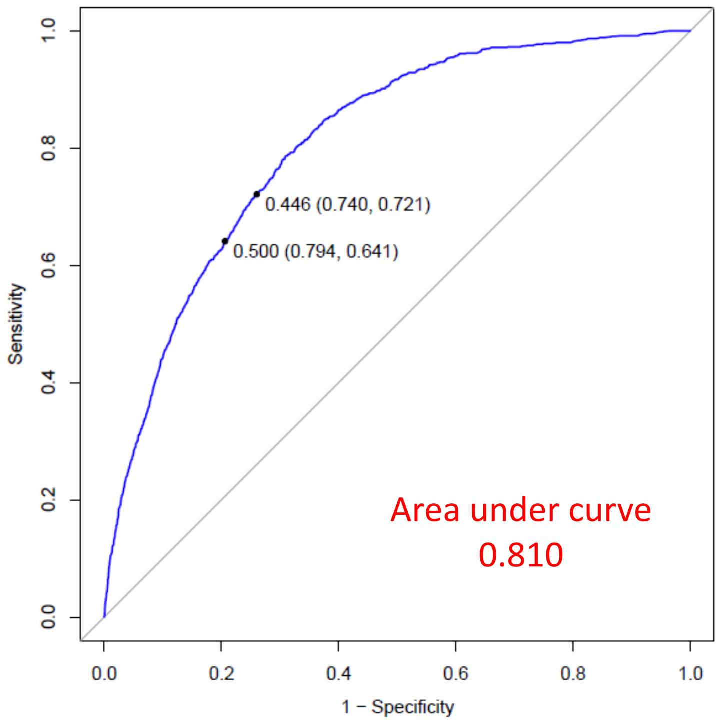
Training Data Results

The final analysis on the training data set incorporated the top 10 predictor variables that were suggested by the random forest model.

	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

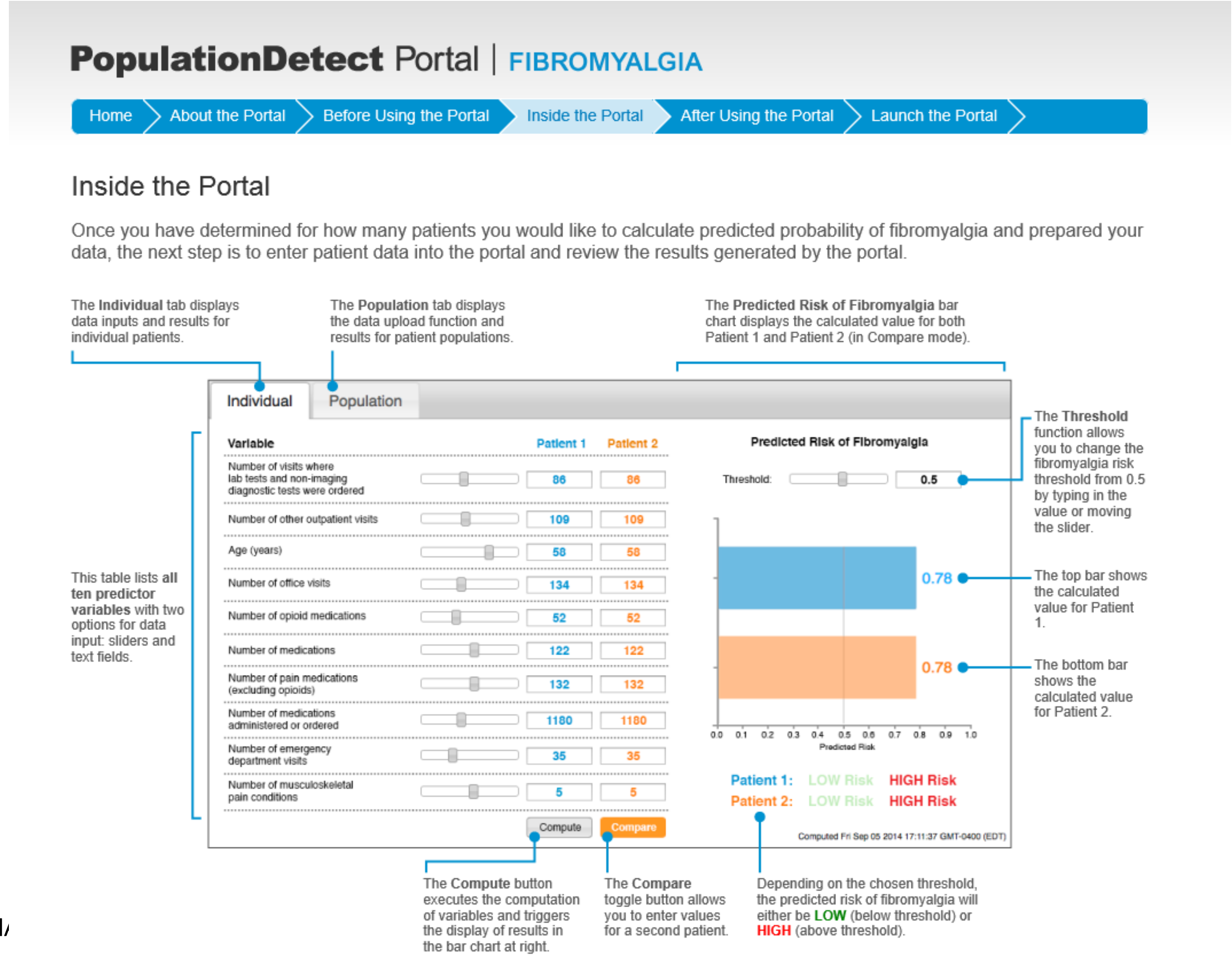



```
CUT OFF 0.5

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 Implementation

- 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 practical application for prediction purposes
- Developed a WEB PORTAL with an easy to use and understand interface

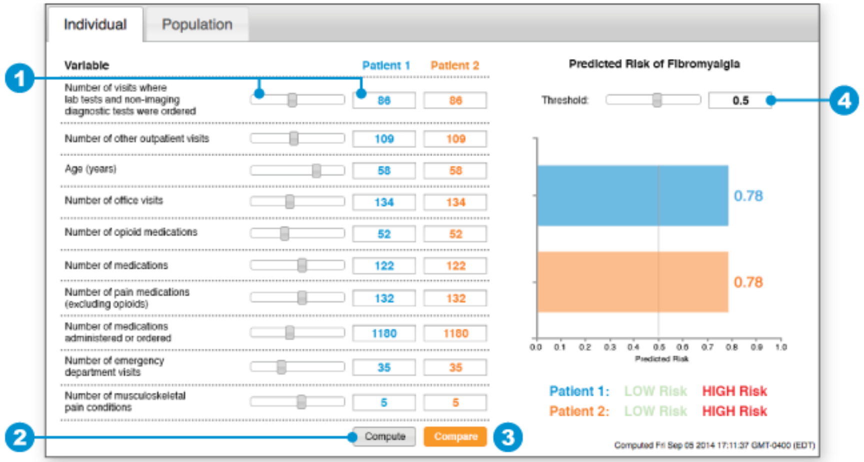
A WEB Interface



Individual Patient Workflow

Individual Patient Workflow

Individual Patient Workflow



When looking at an **individual** patient...

1. Enter the patient's values for each variable using either the sliders or the text fields.

2. Click the "Compute" button. The patient's predicted risk of fibromyalgia will be shown.

3. **Optional:** Click the "Compare" button to enter a second set of values for comparison. Adjust the patient's values using either the sliders or the text fields in the "Patient 2" column. Click "Compute" to view the second patient's predicted risk of fibromyalgia.

4. **Optional:** Use the "Threshold" slider or text field to adjust the threshold between "LOW Risk" and "HIGH Risk."

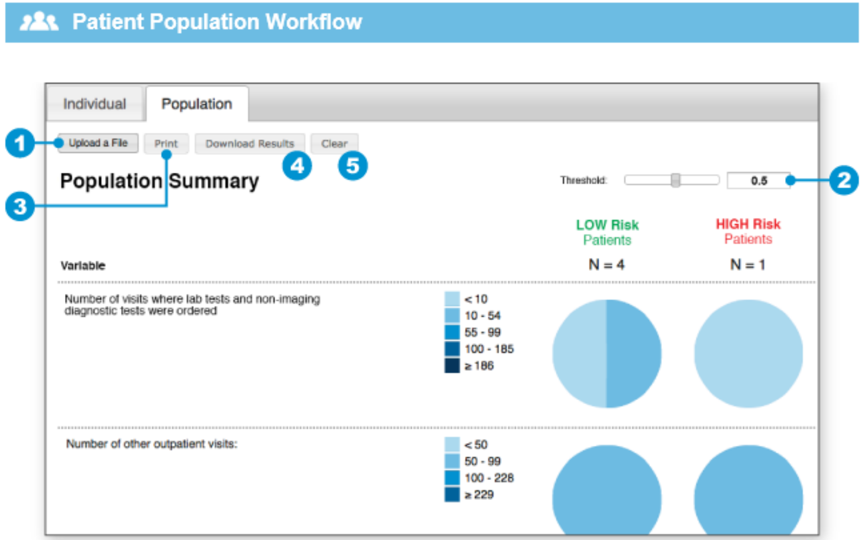
Using C5.0 Rules to Implement Model

C5.0 RULES	VALUE
1. Number of office visits	≤ 9
2. Number of other outpatient visits	≥ 1
3. Number of visits where diagnostics and/or lab tests were ordered	≥ 1
4. Number of emergency department visits	≥ 1
5. Number of opioid prescriptions	≥ 1
6. Number of total pain medications (excluding opioids)	≥ 3
7. Number of total medications administered or ordered	≤ 3
8. Number of musculoskeletal pain conditions	≥ 1

1. Confirm data availability for each variable – to use this method, data must be available in the EHR system for these 8 specific variables. Prorate variables to a yearly rate.

2. Generate a pursuit list: use the full list of 8 rules above to generate a pursuit list in your EHR system.

Patient 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.

FM Project Update

- Manuscript: “Electronic medical record data to identify variables associated with a fibromyalgia diagnosis: importance of health care resource utilization” accepted for publication in Journal of Pain Research
- Manuscript: “Identification of a Potential Fibromyalgia Diagnosis Using Random Forest Modeling Applied to Electronic Medical Records” accepted to Journal of Pain Research (ADHERING TO THE TRIPOD STATEMENT)
- Specialists training sessions conducted. Interaction with payers/providers ongoing
  - Sample Questions regarding model:
    - I increase age but predicted probability of FM does not increase?
    - What if data available on only 9 of 10 variables?
    - Can I start with an enriched population?
    - Can I add variables relevant to A1C?

Last Slide