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

PFIZER

Customer Input

Field Input

FM/pDPN Variables Team
- Marketing
- O&E
- Statistics
- Medical
- Field
- Potential external partners

FM Methodology
- Database
- Subjects
- Model
- Patient Variables

PREDICTIVE MODELING

VARIABLES

FM

x y

10 FM Predictor Variables

RESOURCES

MG/IDS

Leadership &

Data Analyst

FIELD COLLEAGUES

KAM + MOS

Field Training

DIRECT DATA UPLOAD

CUSTOMERS

Pursuit List

Identification of

Undiagnosed Patients

OUTCOME

HCP

Patient

appropriate Patient

Screening & Diagnosis

Medical Community

Paper in Medical Journal
(O&E and Medical)

FM Web Portal

Field Input

Pursuit List

Identification of Undiagnosed Patients

HCP

Patient

Appropriate Patient Screening & Diagnosis

Medical Community

Paper in Medical Journal
(O&E and Medical)

FM Web Portal

Field Input

Pursuit List

Identification of Undiagnosed Patients

HCP

Patient

Appropriate Patient Screening & Diagnosis

Medical Community

Paper in Medical Journal
(O&E and Medical)
• 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
“While working toward this vision of connected care, data are already informing how conditions are diagnosed and managed today. For example, people with the chronic 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 Patient Journey

PRE-DIAGNOSIS

Seek Help
Visit Health Care Provider or Specialist
Receive Incorrect Care
Get Incorrect or No Diagnosis

Costly Cycle
On average, it takes 5 years to diagnose fibromyalgia.²

Don’t Seek Help
Go Unidentified

100 million American adults suffer from chronic pain.¹

DIAGNOSIS

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

Get Symptoms Noticed

Symptoms linked to fibromyalgia may include:
- Uncontrolled pain
- Co-morbid anxiety and/or depression
- Irritable bowel syndrome (IBS)
- Chronic fatigue

Get Screened and Diagnosed

Over 5 million people suffer from fibromyalgia,² but only 36% of fibromyalgia patients are correctly diagnosed.⁴

MANAGEMENT

Fibromyalgia

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

Manage Pain

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

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Method</th>
<th>Class Imbalance</th>
<th>mtry</th>
<th>CV</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Random Forest with 1500 bootstraps</td>
<td>Internal down sampling</td>
<td>13</td>
<td>10-fold repeated 5 times</td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Cut</th>
<th>ROC</th>
<th>Sen</th>
<th>Spe</th>
</tr>
</thead>
<tbody>
<tr>
<td>DS:</td>
<td>0.5</td>
<td>0.824</td>
<td>0.687</td>
</tr>
<tr>
<td>0.446</td>
<td>0.824</td>
<td>0.757</td>
<td>0.741</td>
</tr>
</tbody>
</table>
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 : 0.998740
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
Inside the Portal

Once you have determined for how many patients you would like to calculate predicted probability of fibromyalgia and prepared your data, the next step is to enter patient data into the portal and review the results generated by the portal.

The Individual tab displays data inputs and results for individual patients.

The Population tab displays the data uploaded function and results for patient populations.

The Predicted Risk of Fibromyalgia bar chart displays the calculated value for both Patient 1 and Patient 2 (in Compare mode).

This table lists all ten predictor variables with two options for data input: sliders and text fields.
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.
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.
### 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

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

<table>
<thead>
<tr>
<th>Method</th>
<th>Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>CV</td>
</tr>
<tr>
<td>RPart</td>
<td>10-fold repeated 5 times</td>
</tr>
<tr>
<td>CTree</td>
<td></td>
</tr>
<tr>
<td>k-NN</td>
<td></td>
</tr>
<tr>
<td>RF</td>
<td></td>
</tr>
<tr>
<td>GBM</td>
<td></td>
</tr>
<tr>
<td>SVM</td>
<td></td>
</tr>
<tr>
<td>NB</td>
<td></td>
</tr>
</tbody>
</table>
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.

<table>
<thead>
<tr>
<th>Potential Predictor</th>
<th>ROC</th>
<th>RPart</th>
<th>PLS</th>
<th>RF</th>
<th>GBM</th>
<th>Average Importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pain change from week 3</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
<td>100.0</td>
</tr>
<tr>
<td>Pain change from week 1</td>
<td>76.1</td>
<td>53.4</td>
<td>71.9</td>
<td>36.1</td>
<td>3.1</td>
<td>48.1</td>
</tr>
<tr>
<td>Baseline NRS pain score</td>
<td>0.7</td>
<td>24.0</td>
<td>7.2</td>
<td>19.1</td>
<td>19.7</td>
<td>14.1</td>
</tr>
<tr>
<td>Depression</td>
<td>19.3</td>
<td>15.6</td>
<td>27.7</td>
<td>4.9</td>
<td>1.9</td>
<td>13.9</td>
</tr>
<tr>
<td>Pregabalin as monotherapy</td>
<td>21.4</td>
<td>7.3</td>
<td>25.4</td>
<td>3.9</td>
<td>0.6</td>
<td>11.7</td>
</tr>
</tbody>
</table>
A Spectrum of Predictive Models Applied to an observational data on Neuropathic Pain (NeP)

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.

<table>
<thead>
<tr>
<th>Method</th>
<th>Accuracy (95% CI)</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDA</td>
<td>0.89 (0.88–0.90)</td>
</tr>
<tr>
<td>RPart</td>
<td>0.84 (0.83–0.85)</td>
</tr>
<tr>
<td>CTree</td>
<td>0.83 (0.82–0.85)</td>
</tr>
<tr>
<td>k-NN</td>
<td>0.84 (0.83–0.85)</td>
</tr>
<tr>
<td>RF</td>
<td>0.88 (0.87–0.89)</td>
</tr>
<tr>
<td>GBM</td>
<td>0.88 (0.87–0.89)</td>
</tr>
<tr>
<td>SVM</td>
<td>0.86 (0.85–0.87)</td>
</tr>
<tr>
<td>NB</td>
<td>0.84 (0.82–0.85)</td>
</tr>
</tbody>
</table>
• 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.
Baseline datasets identified…

- to enable independent testing of impacts of each of these challenges on model performance

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Total Patients</th>
<th>Controls</th>
<th>Cases</th>
<th>Prevalence</th>
<th>Label Noise</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED Visit2</td>
<td>720,621</td>
<td>352,984</td>
<td>367,637</td>
<td>51%</td>
<td>0%</td>
</tr>
<tr>
<td>ED Visit</td>
<td>392,204</td>
<td>352,984</td>
<td>39,220</td>
<td>10%</td>
<td>0%</td>
</tr>
<tr>
<td>ED Visit</td>
<td>392,204</td>
<td>352,984</td>
<td>39,220</td>
<td>10%</td>
<td>10%</td>
</tr>
<tr>
<td>ED Visit</td>
<td>392,204</td>
<td>352,984</td>
<td>39,220</td>
<td>10%</td>
<td>20%</td>
</tr>
<tr>
<td>ED Visit</td>
<td>392,204</td>
<td>352,984</td>
<td>39,220</td>
<td>10%</td>
<td>30%</td>
</tr>
</tbody>
</table>

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

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).

<table>
<thead>
<tr>
<th>Tuning Parameters</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Number of hidden layers</strong></td>
</tr>
<tr>
<td><strong>Neurons or cells in each hidden layer</strong></td>
</tr>
<tr>
<td><strong>Learning rate</strong></td>
</tr>
<tr>
<td><strong>Drop-out rate at each layer</strong></td>
</tr>
<tr>
<td><strong>Early Stopping</strong></td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Prevalence</th>
<th>Label Noise</th>
<th>Accuracy¹</th>
<th>Precision¹</th>
<th>Recall¹</th>
<th>F1 Score¹</th>
<th>AU-ROC</th>
</tr>
</thead>
<tbody>
<tr>
<td>51%</td>
<td>~0%</td>
<td>0.7227</td>
<td>0.7688</td>
<td>0.6526</td>
<td>0.7060</td>
<td>0.7942</td>
</tr>
<tr>
<td>50% Cases oversampled from 10%</td>
<td>~0%</td>
<td>0.7237</td>
<td>0.7682</td>
<td>0.6566</td>
<td>0.7080</td>
<td>0.7965</td>
</tr>
<tr>
<td>50% Cases oversampled from 10%</td>
<td>~10%</td>
<td>0.7224</td>
<td>0.7678</td>
<td>0.6535</td>
<td>0.7061</td>
<td>0.7962</td>
</tr>
<tr>
<td>50% Cases oversampled from 10%</td>
<td>~20%</td>
<td>0.7215</td>
<td>0.7673</td>
<td>0.6518</td>
<td>0.7049</td>
<td>0.7951</td>
</tr>
<tr>
<td>50% Cases oversampled from 10%</td>
<td>~30%</td>
<td>0.7216</td>
<td>0.7653</td>
<td>0.6553</td>
<td>0.7060</td>
<td>0.7942</td>
</tr>
</tbody>
</table>

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


• Kuhn and Johnson (2012) Applied Predictive Modeling
Q & A time
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

Step 1: Create Multiple Data Sets
- Randomly select a subset of variables from original data

Step 2: Build Multiple Classifiers
- Build trees

Step 3: Combine Classifiers

Majority wins from this ensemble


Down sampling at each bootstrap sample so that rates of FM and no FM are balanced
Some Remedies for Class Imbalance: Up Sampling, Down Sampling, & SMOTE

Kuhn and Johnson (2012)

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 class.