Johns Hopkins Precision Medicine Analytics Platform (PMAP)

Diana Gumas, M.S.
Senior Director of Clinical Research IT, Institute for Clinical and Translational Research | Instructor, Division of Health Sciences Informatics | Johns Hopkins University

DGUMAS1@JHMI.EDU

Acknowledgement: Thank you to Paul Nagy for many of the slides
What is Precision Medicine at Johns Hopkins

The inHealth Precision Medicine Initiative at Johns Hopkins is bringing together revolutions in measurement, connectivity, and data science to enable the discovery of clinically-relevant and biologically-anchored subgroups at scale.

We know patients with the same diagnosis have different trajectories and responses to treatment.

Experienced clinicians see patterns in subgroups of patients that are similar in some way. They can adjust treatment based on what has worked on similar patients in the past, resulting in better outcomes.

We want to enable you to find and validate those subgroups faster, by providing more data, more efficiently, along with the tools to analyze it all. We are giving you the tools to find the underlying mechanisms. And we are helping you put those discoveries back into clinical care quickly.
Precision Medicine Initiative at Johns Hopkins
2017
- Prostate Cancer
- Multiple Sclerosis

2018
- Neurofibromatosis
- Myositis
- Scleroderma
- Bladder Cancer
- Pancreatic Cancer
- Heart Arrhythmias
- Population Health State

2019
- Lung Cancer
- COPD
- Precision Rehabilitation
- Breast Cancer
- Alzheimer’s
- Chronic Renal Failure
- Loeys-Dietz Syndrome
- Neurology Critical Care

Data Analysis 600.720
PMAPE Cookbook
PMCoE Lite

Ground Teams
PMCoE “In a Box”

6-10 new
Precision Medicine
Centers of Excellence (PMCoE) every year

50+ by 2024
# 4 Challenges for Clinical Research

<table>
<thead>
<tr>
<th></th>
<th><strong>Data Access:</strong> Construct an environment with data security protecting patient information recognizing complete de-identification is difficult</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>Cross Modality:</strong> Combine access to EMR, medical imaging, genomics, and physiological monitoring data in one location.</td>
</tr>
<tr>
<td>3</td>
<td><strong>Data Science:</strong> Equip clinical researchers with the training and tools of data science.</td>
</tr>
<tr>
<td>4</td>
<td><strong>Delivery:</strong> Bring discoveries into clinical care through context integrated clinical decision support platform.</td>
</tr>
</tbody>
</table>
PMAP from the 10,000 foot level

DATA SOURCES
- Pathology
- Epic
- Radiology
- Study Databases from REDCap, Access
- Future: Legacy EMR DB, External Data

DISCOVERY
- DATA COMMONS
  Hosts all data coming into platform from various, disparate data sources
- RESEARCH ENVIRONMENT
  Secure user environment with aggregated data sets approved for user – allows advanced analytics and other tools to be applied to data

TREATMENT PLATFORM
- Shared Algorithms, Analytics, Clinical Care Impact, Products
- Commercialize

Reusable Pipelines
IRB Approvals
Validation/Promotion Process
Feedback
Data on PMAP

Microsoft Azure Cloud

Hortonworks Hadoop Data Platform
<table>
<thead>
<tr>
<th></th>
<th><strong>What does PMAP do for me?</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td><strong>IRB application time:</strong> PMAP is approved by the Data Trust and the IRB as a secure research environment.</td>
</tr>
<tr>
<td>2</td>
<td><strong>Data Access:</strong> One stop shop for access to EMR, medical imaging, genomics, and physiological monitoring data.</td>
</tr>
<tr>
<td>3</td>
<td><strong>Analysis tools:</strong> PMAP provides state of the art tools and resources needed to analyze the data.</td>
</tr>
<tr>
<td>4</td>
<td><strong>Team Science:</strong> PMAP helps you to share the data and analysis with your research team.</td>
</tr>
</tbody>
</table>
Jupyter Notebook: Data Science at Scale

Visual comparison with paired plotting in seaborn

Seaborn comes with a very useful pairplot function which will provide a scatter plot of all the fields in a dataset against each other.

Key things to look for:

1. Using python list comprehension we selected only the columns that ended with the string "mean". There are over 30 variables and we wanted to narrow down the field.
2. In the pairplot command we filter the list of columns to those selected to "mean" and color the data according to its diagnosis.

The goal is to look for pairs of variables that can best separate malignant from benign cells.

In [8]:

```python
# Select only the columns with a mean measurement

cols=[x for x in df.columns if x[-4:]=="mean"]

tsns.pairplot(df,vars=cols,hue='diagnosis')

dataframe.show()
```

- Web based computational scientific notebooks.
- Open standard file format for reproducible research.
- Supports multiple languages (R, Python)
- Can integrate with high performance compute clusters
Johns Hopkins Precision Medicine Analytics Platform (PMAP) Cookbook

About the PMAP Cookbook

Table of Contents

Before you start cooking

- Do you need an IRB to access PMAP
- What type of protocol should I submit for my research use of PMAP?
- Do I have to get Data Trust approval to use PMAP?
- How can I tell what data are available in PMAP?

Applications on PMAP

- Crunchr collaborative data science platform
  - How to access the PMAP Cookbook on Crunchr
  - Steps to using Crunchr for collaborative research
- Introduction to Jupyter notebooks

Working with EMR Data

- Connecting to the PMAP database server
  - Requesting access to the asthma dataset
- Intro to SQL for mining EMR data
- Asthma cohort discovery with regular expressions and set operators
- Exploratory data analysis and visualization of EMR data
- Data quality assessment example
- Statistical testing EMR Asthma example

Conducting NLP Projects

- Tokenizing a corpus of text
- Matching text with Regular Expression
- n-grams and term weighting
- Document classification

Working with Medical Images

- Working with DICOM image files
- Radiomics Example
- High Performance Compute (HPC) clusters for Image Analysis
- Hello World Chest X-Ray Deep learning example

Programming Resources

- Python Data Science Handbook (O'Reilly Media)
  - Python programming
  - Data Manipulation with Pandas
  - Data Visualization with Matplotlib
  - Introduction to Numpy
  - Machine Learning with Scikit

http://pm.jh.edu/cookbook/
PMP Imaging Functionality

- **Archive** – Federated by VNA
- **Index** - Imaging metadata (DICOM)
- **Search** - Users can query DICOM with Hive’s SQL-like language
- **Access** – Retrieve images to their environment
- **Annotation** – UI to develop features for supervised learning.
- **Processing** – Docker containers for registration, segmentation
- **Analysis** – Jupyter/SciServer
- **Compute** - Deep learning GPU for Tensorflow/Keras
import matplotlib.pyplot as plt

print(history.history.keys())

plt.figure()
plt.plot(history.history['acc'], 'orange', label='Training accuracy')
plt.plot(history.history['val_acc'], 'blue', label='Validation accuracy')
plt.plot(history.history['loss'], 'red', label='Training loss')
plt.plot(history.history['val_loss'], 'green', label='Validation loss')
plt.legend()
plt.show()

dict_keys(['val_loss', 'acc', 'val_acc', 'loss'])
Natural Language Processing

```python
# Scoring metrics we're interested in
scoring = {'acc': 'accuracy',
           'precision': 'precision',
           'recall': 'recall',
           'f1': 'f1'}

# Want to initialize the random number generator with the same seed for repeatable experiment:
random_state = 198273

# Our preprocessing steps

def preprocess(X):
    return preprocess_string(X, filters=[lambda x: x.lower()])

preprocess_pipeline = Pipeline([
    ('vect', CountVectorizer(analyzer=preprocess)),
    ('tfidf', TfidfTransformer()),
])

features = preprocess_pipeline.fit_transform(text)

for clf in [RandomForestClassifier(),
            MultinomialNB(),
            SVC(kernel='Linear', class_weight='balanced'),
            DecisionTreeClassifier(criterion='entropy', class_weight='balanced')]:
    scores = cross_val_score(clf, features, labels, scoring=scoring, cv=5, return_train_scores=False)
    print("Classifier:", clf)
    # print(scores)
    print("Accuracy:" , scores['test_acc'].mean())
    print("Precision:" , scores['test_precision'].mean())
    print("Recall:" , scores['test_recall'].mean())
    print("F1:" , scores['test_f1'].mean())
    print()
```

---

**Example:**

1. **PROSTATE, LEFT APEX (BIPSY): ADENOCARCINOMA OF THE PROSTATE, GLEASON GRADE 3+3=6 INVOLVING ONE (1) OF TWO (2) CORES (30%).**

2. **E2 LEFT MID: MINUTE FOCUS OF ADENOCARCINOMA OF THE PROSTATE, GLEASON GRADE 3+3=6 INVOLVING ONE OF ONE BISECTED CORE.**

3. **PROSTATIC ADENOCARCINOMA, GLEASON SCORE 3+3=6 (GRADE GROUP 1) INVOLVING 20% OF ONE (1) CORE. SEE NOTE.**

4. **LEFT LATERAL APEX: SMALL FOCUS OF PROSTATIC ADENOCARCINOMA, GLEASON SCORE 3+3=6, INVOLVING LESS THAN 5% OF ONE (1) CORE OF ONE CORE.**

5. **MINUTE FOCUS OF ADENOCARCINOMA OF THE PROSTATE, GLEASON GRADE 3+3=6 INVOLVING ONE (1) OF TWO CORES.**

6. **RIGHT APEX: SMALL FOCUS OF PROSTATIC ADENOCARCINOMA, GLEASON SCORE 3+3=6, INVOLVING ONE (1) CORE.**

7. **RIGHT BASE: ADENOCARCINOMA OF THE PROSTATE, GLEASON GRADE 3+3=6 INVOLVING 3 OF 3 CORES (80%, 80%, 50%). PERINEURAL INVASION IDENTIFIED.
De-identified Epic Dataset

<table>
<thead>
<tr>
<th>Table Name</th>
<th># of Records</th>
<th># of Columns</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encounter</td>
<td>690,183</td>
<td>4</td>
<td>36.57 MB</td>
</tr>
<tr>
<td>Labs</td>
<td>3,486,911</td>
<td>12</td>
<td>403.41 MB</td>
</tr>
<tr>
<td>Meds</td>
<td>5,926,733</td>
<td>9</td>
<td>1.86 GB</td>
</tr>
<tr>
<td>Patients</td>
<td>60,676</td>
<td>5</td>
<td>7.18 MB</td>
</tr>
<tr>
<td>ProblemList</td>
<td>115,162</td>
<td>4</td>
<td>16.99 MB</td>
</tr>
<tr>
<td>Procedure</td>
<td>6,449</td>
<td>6</td>
<td>0.727 MB</td>
</tr>
<tr>
<td>Symptom</td>
<td>28,056</td>
<td>5</td>
<td>1.711 MB</td>
</tr>
<tr>
<td>Vitals_BP</td>
<td>390,181</td>
<td>7</td>
<td>30.58 MB</td>
</tr>
<tr>
<td>Vitals_Height</td>
<td>279,288</td>
<td>7</td>
<td>18.96 MB</td>
</tr>
<tr>
<td>Vitals_Pulse</td>
<td>388,450</td>
<td>7</td>
<td>65.92 MB</td>
</tr>
<tr>
<td>Vitals_Respiration</td>
<td>251,166</td>
<td>7</td>
<td>18.18 MB</td>
</tr>
<tr>
<td>Vitals_Temperature</td>
<td>314,571</td>
<td>7</td>
<td>21.87 MB</td>
</tr>
<tr>
<td>Vitals_Weight</td>
<td>352,553</td>
<td>7</td>
<td>24 MB</td>
</tr>
<tr>
<td><strong>TOTAL</strong></td>
<td><strong>12,290,379</strong></td>
<td><strong>87</strong></td>
<td><strong>2.55 GB</strong></td>
</tr>
</tbody>
</table>

60,000 patients
112 M Data Elements
No note fields
Delivery: Longitudinal Visualization of Patient Data

Longitudinal view of medications
Gantt chart view of individual medications

Medical Regimens
Changes in medical regimen are represented by shaded vertical bands.

Patient Events
Patient Encounters, Relapses, MRIs…

Clinical progression data series
Plottable data items, lab values, test results are displayed as line series.
## Ways to work with PMAP today

<table>
<thead>
<tr>
<th></th>
<th>Free analytics and compute platforms: Work with our de-identified dataset. SAFE windows client. Crunchr pooled data science cluster.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><strong>Training opportunities.</strong> Cookbook, CAMP, and clinical data analysis programs.</td>
</tr>
<tr>
<td>3</td>
<td><strong>Static data requests.</strong> Use PMAP as part of your IRB request and work with CCDA ($/hr) and extract and upload data.</td>
</tr>
<tr>
<td>4</td>
<td><strong>Create a prospective registry (Annual Fee).</strong> Apply to become a PM Center of Excellence (funded) Use PMAP to apply for grants.</td>
</tr>
</tbody>
</table>
Recognition and Thanks to colleagues at Hopkins

- Paul Nagy, PhD
- Ken Pienta, MD
- Aalok Shah
- Dwight Raum
- John Scott
- Alan Coltri
- Corban Rivera, PhD
- Antony Rosen, MD
- Scott Zeger, PhD
- Chris Chute, MD, PhD
- Pat Ostendarp
- Phil Gianuzzi
- Geoff Osier
- Ken Harkness
- Brant Chee, PhD
- Alex Baras, MD
- Jerry Prince, PhD
- Bal Carter, MD
- Ellen Mowry, MD
- Peter Calibresi, MD
- Rai Winslow, PhD
- Lee Watkins, PhD
- Masoud Rouhzadeh, PhD
- Benjamin Smith
- Peter Greene, MD
- Steven Miller, MD
- Bonnie Woods
- Thomas Grader-Beck MD
- Amy Hushen
- Katie Patras
- David Li
- Chris Doyle
- Jasmine McNeil
- Will Roncal, PhD
- Alex Oshmansky
- Anna Duerr
- Caitlin Bishop, PhD
- Michael Johnson, MD
For more information go to the Precision Medicine Portal

http://pm.jh.edu

For the latest updates follow on Twitter @hopkinsinhealth