Genetic risk modeling using machine learning to predict radiotherapy complications and identify key biological correlates

Jung Hun Oh
Assistant Attending
Department of Medical Physics,
Memorial Sloan Kettering Cancer Center

Acknowledgement

GWAS Study
- Sangkyu Lee, MSK
- Joseph Deasy, MSK
- Barry Rosenstein, Mount Sinai School of Medicine
- Sarah Kema, University of Rochester Medical Center
- Harry Ostrer, Albert Einstein College of Medicine

Radiomics
- Evangelia Katsoulakis, MSK
- Yao Yu, MSK
- Aditya P. Apte, MSK
- Nancy Y. Lee, MSK
- Nadeem Riaz, MSK
- Vaios Hatzoglou, MSK
- Jonine Bernstein, MSK
- Xiaolin Liang, MSK
- Meghan Woods, MSK
- Anne Reiner, MSK

Part 1. Genome-wide association studies
Background

➢ Our goal in GWAS is to predict how the risk of radiation toxicity varies between patients, based on germ line genome characteristics
➢ Previous single-SNP models suffer from multiple-testing correction due to a large number of SNPs being evaluated
➢ Important SNPs may fail to achieve genome-wide significance
➢ Therefore, we have taken a many-SNP approach to developing predictive models, using machine learning methods

Single nucleotide polymorphisms (SNPs)

https://www.tubascan.eu

Genome-wide association studies (GWAS)

Patient 1 ...CAAGGTA...
Patient 2 ...CAATGTA...
Patient 3 ...CAATGTA...
Patient 4 ...CAAGGTA...

Single Nucleotide Polymorphism (SNP) is genetic variation at one location in a DNA sequence.

Genome-Wide Association Studies (GWAS) find associations between a disease and such variations across the whole genome.
Coding

➢ Wild type homozygous: 0
➢ Heterozygous: 1
➢ Mutant homozygous: 2

➢ Coded as the number of rare alleles

Population structure


Statistical analysis

Manhattan Plot

❖ Genome-wide significance level = 5 x 10^-8
Filtering

-\log(p\text{-value})

- \text{p} = 5 \times 10^{-8}

- \text{p} = 1 \times 10^{-1}

Real biomarkers
Non-biomarkers

Preconditioning

Training data
Univariate Analysis

V1
V2

AUC

SNPs

Top 1
Top 2
Top 3
Top N

Predicted outcomes

Best AUC

0.89

preconditioned outcomes

SNP importance

SNP 1
SNP 2

OOB error = \frac{1}{\text{sample size}} \sum \text{error}

Importance of SNP 1 = \frac{\text{Rand. error} - \text{OOB error}}{\text{Rand. error}}

Randomization

SNP 1

SNP 2

SNP 3

Training data
Bootstrapping data

Data not used in tree = 36.8%

OOB error = 0.88

– 0.80 = 0.08

Error = 0.38

– 0.18 = 0.20

Error = 0.82
A subset of features

Gene ontology analysis

Preconditioning Random Forest Regression (PRFR)

Oh et al., 2017. Scientific Reports
Data

➢ 368 patients with prostate cancer
  - DNA was genotyped using Affymetrix genome wide array (v6.0)
➢ Quality control
  - Missing rate > 5% of samples
  - MAF < 5%
  - Hardy-Weinberg equilibrium (p-value < 10^-5)
  - 613,496 SNPs remained

Rectal bleeding

Oh et al., 2017. Scientific Reports

Data preprocessing

➢ Outcome: rectal bleeding
  - RTOG ≤ 1 (coded 0) vs RTOG ≥ 2 (coded 1)
➢ Data split: rectal bleeding
  - Training dataset
    - 243 samples
    - 49 events
    - 749 SNPs (p< 0.001; Chi-square test)
  - Validation dataset
    - 122 samples
    - 25 events
➢ 5-fold cross validation with 100 iterations
### Performance

1. **Regulation of ion transport**
   - CACNA1D, CCL13, DPP6, GCK, GNB4, GPR61, HOMER1, JLP, JLP1, TCP1
   - CACNA1D, CCL13, DPP6, GCK, GNB4, GPR61, HOMER1, JLP, JLP1, TCP1

2. **Regulation of metal ion transport**
   - CACNA1D, CCL13, DPP6, GCK, GNB4, GPR61, HOMER1, JLP, JLP1, TCP1

3. **Regulation of ion transmembrane transporter activity**
   - CACNA1D, GNB4, HOMER1, JLP, JLP1, TCP1

4. **Regulation of ion transmembrane transport**
   - CACNA1D, CCL13, DPP6, GCK, GNB4, GPR61, HOMER1, JLP, JLP1, TCP1

5. **Regulation of transmembrane transporter activity**
   - CACNA1D, GNB4, HOMER1, JLP, JLP1, TCP1

### Biological analysis

1. **GO Processes/Genes**
   - Regulation of ion transport

2. **Epidermal Growth Factor Partially Restores Colonic Ion Transport Responses in Mouse Models of Chronic Colitis**
   - Declan F. McCole, Gerhard Roessler, Neogi Yamazaki, and Mark E. Barrett

3. **GO Processes/Genes**
   - Regulation of transmembrane transporter activity
Data preprocessing

- Outcome: erectile dysfunction
  - SHIM ≤ 7 (coded 1) vs SHIM ≥ 16 (coded 0)
- Data split
  - Training dataset
    - 157 samples
    - 88 events
    - 367 SNPs (p < 0.001; Chi-square test)
  - Validation dataset
    - 79 samples
    - 45 events
Performance

Biological analysis

<table>
<thead>
<tr>
<th>Ranking</th>
<th>GO Processes</th>
<th>FDR</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>negative regulation of heart contraction</td>
<td>8.276E-10</td>
</tr>
<tr>
<td>2</td>
<td>negative regulation of blood circulation</td>
<td>2.180E-08</td>
</tr>
<tr>
<td>3</td>
<td>neutrophil chemotaxis</td>
<td>5.026E-08</td>
</tr>
<tr>
<td>4</td>
<td>neutrophil migration</td>
<td>5.883E-08</td>
</tr>
<tr>
<td>5</td>
<td>granulocyte chemotaxis</td>
<td>9.684E-08</td>
</tr>
<tr>
<td>6</td>
<td>granulocyte migration</td>
<td>1.300E-07</td>
</tr>
<tr>
<td>7</td>
<td>positive regulation of locomotion</td>
<td>2.831E-07</td>
</tr>
<tr>
<td>8</td>
<td>regulation of muscle system process</td>
<td>5.510E-07</td>
</tr>
<tr>
<td>9</td>
<td>regulation of muscle contraction</td>
<td>5.510E-07</td>
</tr>
<tr>
<td>10</td>
<td>positive regulation of cell migration</td>
<td>8.960E-07</td>
</tr>
</tbody>
</table>
Genitourinary Toxicity

Lee et al., 2018. Int J Rad Oncol Biol Phys

<table>
<thead>
<tr>
<th>Symptom Category*</th>
<th>Symptom Name</th>
<th>Training Set Size</th>
<th>Testing Set Size</th>
<th>Event Rate (%)</th>
<th>Modeled?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Irritative (Storage)</td>
<td>Frequency</td>
<td>119</td>
<td>60</td>
<td>23</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Urgency</td>
<td>161</td>
<td>81</td>
<td>16</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Nocturia</td>
<td>111</td>
<td>56</td>
<td>17</td>
<td>O</td>
</tr>
<tr>
<td>Obstructive ( Voiding)</td>
<td>Intermittency</td>
<td>164</td>
<td>82</td>
<td>10</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Weak Stream</td>
<td>149</td>
<td>75</td>
<td>18</td>
<td>O</td>
</tr>
<tr>
<td></td>
<td>Straining</td>
<td>196</td>
<td>98</td>
<td>5</td>
<td>X</td>
</tr>
<tr>
<td></td>
<td>Incomplete Emptying</td>
<td>168</td>
<td>84</td>
<td>10</td>
<td>X</td>
</tr>
</tbody>
</table>
GU modeling

<table>
<thead>
<tr>
<th>Symptom Name</th>
<th>Training Set Size</th>
<th>Event Rate</th>
<th># SNPs</th>
<th># clinical</th>
<th>SNPs p&lt;0.001</th>
<th>SNPs p&lt;0.05</th>
<th>PRFR Performance</th>
<th>AUC</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency</td>
<td>119</td>
<td>0.23</td>
<td>539</td>
<td>0</td>
<td>0.64</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Urgency</td>
<td>161</td>
<td>0.16</td>
<td>738</td>
<td>0</td>
<td>0.53</td>
<td>0.38</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nocturia</td>
<td>111</td>
<td>0.17</td>
<td>977</td>
<td>1</td>
<td>0.55</td>
<td>0.33</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weak-stream</td>
<td>149</td>
<td>0.18</td>
<td>823</td>
<td>0</td>
<td>0.70</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Gene ontology analysis

Protein-protein network
Summary

➢ We developed a promising method using whole-genome data for deriving predictive risk models for predicting late radiation-induced toxicities
➢ SNP -> Gene -> Pathway analysis
➢ Found that biological correlates are tissue specific
➢ Other studies
  - Secondary contralateral breast cancer
  - Fatigue in breast cancer
  - Weight gain in breast cancer

Part 2. Radiogenomics

Data

➢ Imaging data:
  - Pre-treatment CT scans in head and neck cancer were downloaded from the TCIA
  - 77 CT scans were analyzed

  - Using CERR, 104 radiomic features were evaluated
    - Apte, 2018. Medical Physics
    - Feature stability test
    - Volume dependent features were removed
    - 67 features were analyzed
Data

- **Biological data:**
  - Recurrent gene mutations
    - cBioPortal (https://www.cbioportal.org/)
  - Tumor subtypes
    - Broad Institute FireBrowse (http://firebrowse.org)
  - Immune infiltrates
    - Thorsson, 2018. Immunity
  - HPV status
    - Nulton, 2017. Oncotarget

### Clustering

<table>
<thead>
<tr>
<th>Subsite</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral cavity</td>
<td>15</td>
<td>23</td>
</tr>
<tr>
<td>Larynx</td>
<td>14</td>
<td>14</td>
</tr>
<tr>
<td>Oropharynx</td>
<td>10</td>
<td>1</td>
</tr>
</tbody>
</table>

P-value: 0.0006

<table>
<thead>
<tr>
<th>HPV</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>11</td>
<td>2</td>
</tr>
<tr>
<td>Negative</td>
<td>28</td>
<td>36</td>
</tr>
</tbody>
</table>

P-value: 0.0127
### Representative scans

Cluster 1  
Cluster 2

### CD8 Prediction using Random Forest

<table>
<thead>
<tr>
<th>Status</th>
<th>R²</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>All</td>
<td>0.3</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>HPV+</td>
<td>0.36</td>
<td>0.0405</td>
</tr>
<tr>
<td>HPV-</td>
<td>0.16</td>
<td>0.0012</td>
</tr>
</tbody>
</table>

### HPV status vs CD8 T-cell

P-value = 0.0051
Validation

<table>
<thead>
<tr>
<th>Subsite</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Oral cavity</td>
<td>14</td>
<td>27</td>
</tr>
<tr>
<td>Larynx</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>Oropharynx</td>
<td>27</td>
<td>4</td>
</tr>
<tr>
<td>P-value</td>
<td>1.3x10^-7</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>HPV</th>
<th>Cluster 1</th>
<th>Cluster 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>Positive</td>
<td>24</td>
<td>3</td>
</tr>
<tr>
<td>Negative</td>
<td>17</td>
<td>30</td>
</tr>
<tr>
<td>P-value</td>
<td>4.6x10^-7</td>
<td></td>
</tr>
</tbody>
</table>

- 83 cases (MSKCC)
- Oral cavity: 51, Larynx: 1, Oropharynx: 31

Difference of HPV status

- Sensitivity test
  - Randomly select 5%, 10%, 15%, and 20% of samples in oral cavity tumors
  - Assign them to HPV-positive
  - Iterate 1000 times

  - Prevalence of HPV incidence: 5%
    \[ P=1.4 \times 10^{-5} \text{ (95\% CI: 1.3 \times 10^{-5}-1.5 \times 10^{-5})} \]

Summary

- Found clearly separable radiomic clusters
- The differences in subsite and HPV status between the two radiomic clusters were statistically significant
- Built a machine learning model to predict CD8 T-cell
- Validation using an independent dataset
Thank you