Predicting Cancer Risks Via Multi-Parametric Correlations

Jun Deng, PhD
Professor
Department of Therapeutic Radiology
Yale University School of Medicine

August 1, 2017, Denver, CO, AAPM Annual Meeting

Yale SCHOOL OF MEDICINE

Disclosure

No conflicts of interest to disclose

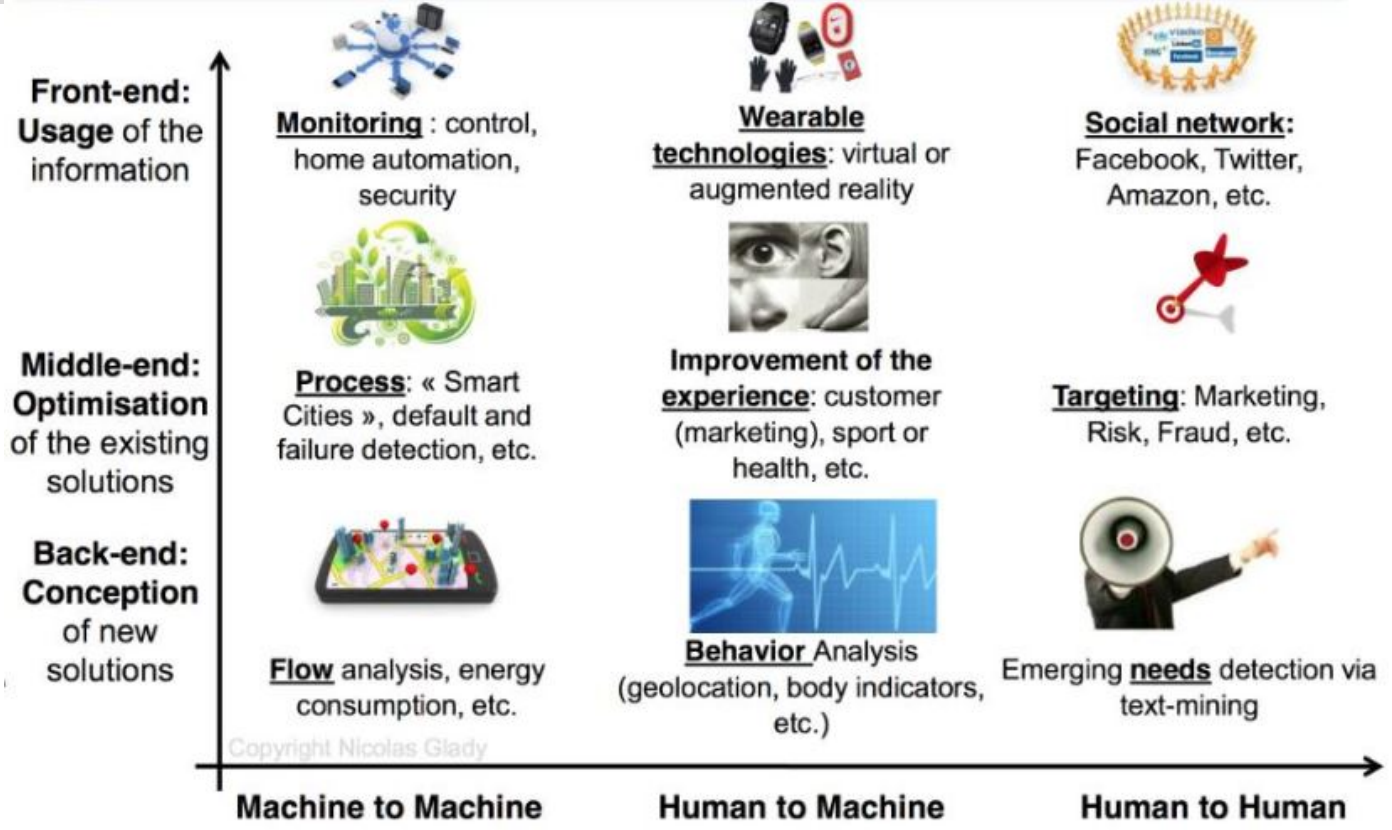
Contents

- Big data basics
- Machine learning 101
- Big data in radiation oncology and applications
- Prostate cancer risk prediction via deep neural network
- Conclusions
- Outlook

We Live in An Ever-Growing Data World

- Over 90% of all the data in the world was created in the past 2 years
- Every 2 days we created as much information as we did from the beginning of time until 2003

Risky? Maybe. But also a good opportunity!



Target Knows and Predicts

Target Knows and Predicts

- Each customer gets an ID, tied to credit card, name, email address, purchase history, and any demographic information
- Analyze historical buying data for all the ladies who have signed up for Target baby registries in the past
- Look for time-purchasing patterns
- Predict what the consumers most likely to buy next time
- Mail out coupons that are most likely to make consumers happy

Target Knows and Predicts

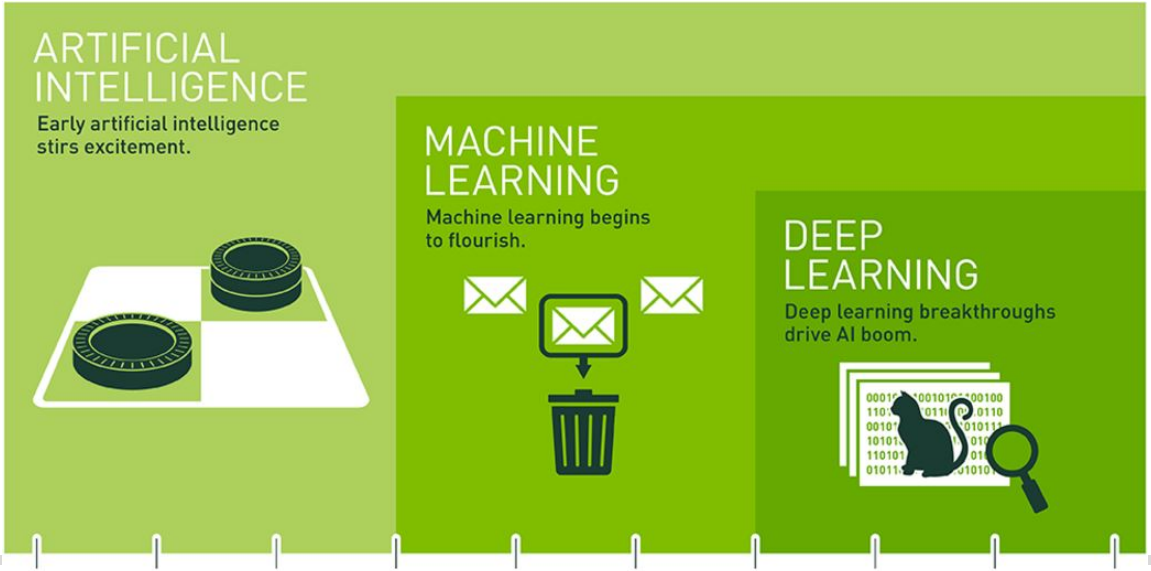
You are what you buy

Big Data Characteristics

- Four V's: **V**olume, **V**ariety, **V**elocity, and **V**eracity
- **Volume**: a large volume of data collected and stored continuously
- **Variety**: structured data in traditional databases, and unstructured text documents, emails, video, audio, notes and financial transactions
- **Velocity**: data is streaming in at unprecedented speed
- **Veracity**: bias, noise and abnormality in data
- What is important in big data analysis is **correlation** not causality

Machine Learning 101

- Artificial Intelligence has exploded since 2015
  - GPUs make parallel processing ever faster, cheaper, and more powerful
  - Big Data pouring in: images, text, transactions, mapping data
- Deep learning seeks to model data, decipher correlations and make decisions



Machine Learning Algorithms

- Information-based machine learning
  - Decision tree
  - Random forest
- Similarity-based machine learning
  - K nearest neighbor (KNN)
- Probability-based machine learning
  - Naïve Bayes
  - Markov chain Monte Carlo
- Error-based machine learning
  - Logistic regression
  - Support vector machines (SVM)
  - Artificial neural networks (ANN)

Machine Learning Algorithms

- Supervised machine learning
  - Decision tree
  - Random forest
  - Logistic regression
  - K nearest neighbor
  - Artificial neural networks
- Unsupervised machine learning
  - Apriori algorithm
  - K-means
- Reinforcement learning
  - Markov Decision Process
  - Deep reinforcement learning (e.g., AlphaGo)

Deep Blue vs Kasparov

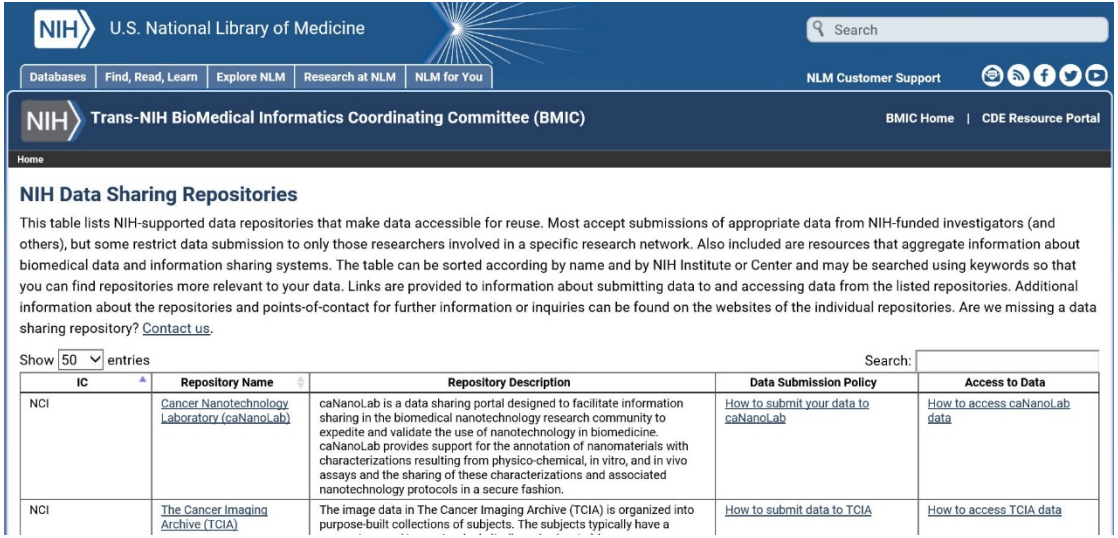
- IBM Deep Blue used a brute force search approach to beat Kasparov in 1997
- Deep Blue goes through all the possible moves to a depth of 6 to 20 moves

AlphaGo vs Lee Sedol & Ke Jie

- There are 10^{170} possible positions in Go, too many to try a brute force search
- Google AlphaGo uses deep reinforcement learning to teach the machine to self-learn from its own moves, improve, and make better moves

Big Data Resource in Cancer and Biomedical Research

- National Cancer Database (NCDB): https://www.facs.org/quality-programs/cancer/ncdb
- NIH Big Data to Knowledge (BD2K): https://bd2kccc.org/
- NIH Data Sharing: https://www.nlm.nih.gov/NIHbmic/nih_data_sharing_repositories.html



Big Data in Radiation Oncology

Table 1 Sizes of genomic data compared to some existing clinical data domains

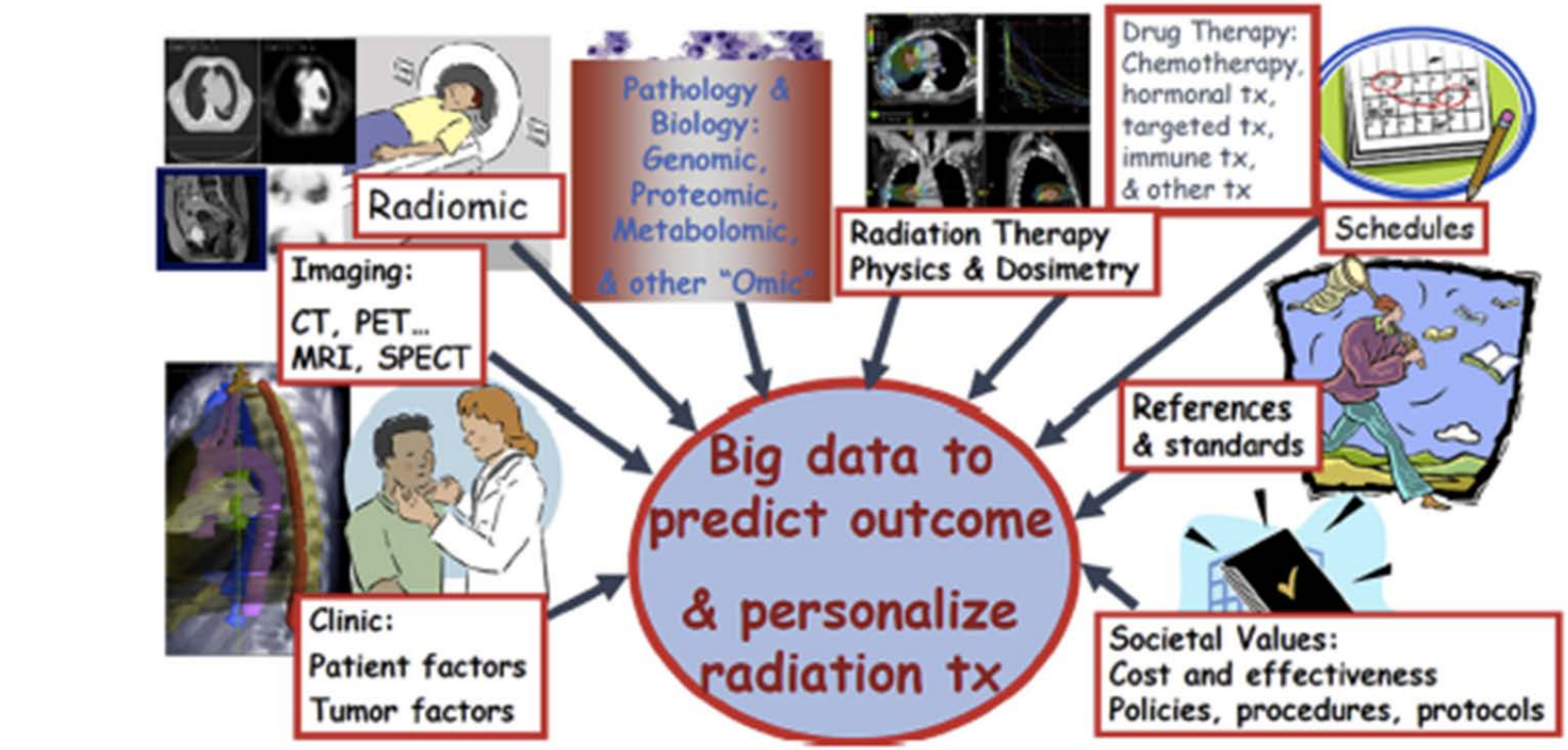
Data type	Data elements	Single patient (average)	Cohort of 1 million patients
Clinical reports	Text	10 MB	10 TB
Laboratory results	Value, units, flag	0.3 MB	0.3 TB
Administrative plus EHR data	Dx, Proc, Rx	2 MB	2 TB
Exome genomic data (variants) (VCF)	Position, type, base(s)	125 MB	125 TB
Imaging data	Multiple image formats	421.9 MB*	421.9 TB*
Total		559.2 MB	559.2 TB
Raw exome genomic data (BAM)	Position, base, quality	5.7 GB	5.7 PB
Grand total		6.3 GB	6.3 PB

Abbreviations: BAM = binary alignment/map; Dx = diagnosis; EB = exabyte (10^{18}); EHR = electronic health record; GB = gigabyte (10^{9}); MB = megabyte (10^{6}); PB = petabyte (10^{15}); Proc = procedure; Rx = prescription; TB = terabyte (10^{12}); VCF = variant call format.

* Imaging data estimate does not represent an average patient but is based on the cancer patient cohort in the Cancer Imaging Archive (13.5 TB of image data for approximately 32,000 cancer patients [data as of April 2015]) (4).

Huser and Cimino, Int J Radiol Oncol Biol Phys 95(3), 890-894, 2016

Tap Big Data in Radiation Oncology



Rosenstein et al, Int J Radiol Oncol Biol Phys 2016

The Question We Try to Answer

- Can we achieve individualized cancer risk prediction via machine learning with big health data?

CDC National Health Interview Survey

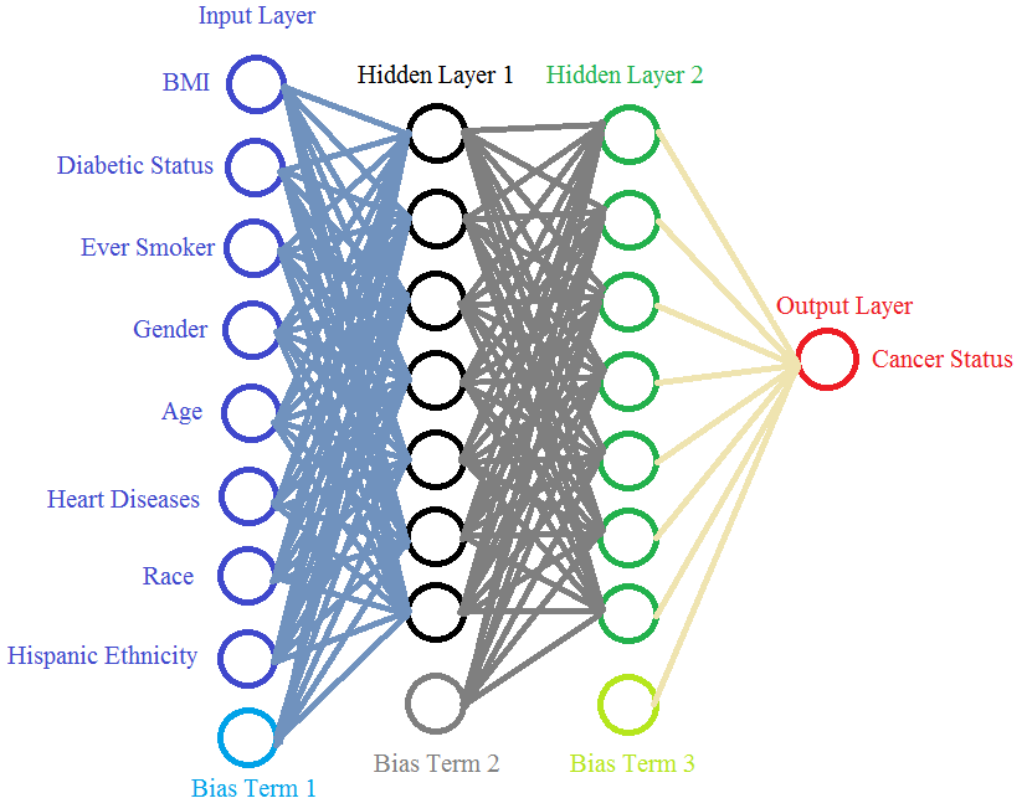
- Publically available CDC data from 1997-2015
- Total observations: 555,183
- Variables of interest:

  Age, Sex, Race, BMI, Smoking, Asthma, Diabetes, Strokes, Hypertension, Family History, Alcohol consumption, Hispanic ethnicity, Cardiovascular Disease, Physical Exercise, Chronic Obstructive Pulmonary Disease (COPD)

Demographics of the Data	Prostate Cancer	Non-Cancer
Average Age	68.94	45.19
Average BMI	27.83	27.56
Percentage That Have Ever Smoked	63.10%	49.02%
Percentage That Have COPD	4.69%	1.74%
Percentage That Have Asthma	8.97%	9.35%
Percentage That Have Diabetes	17.88%	7.89%
Percentage That Have Ever Had a Stroke	7.25%	2.39%
Percentage with Hypertension	60.31%	26.66%
Average Heart Disease Score	13.51%	4.41%
Percentage White	77.24%	79.01%
Percentage African American	19.61%	13.45%
Percentage Native American/Alaska Native	0.48%	0.87%
Percentage Asian	1.72%	5.16%
Percentage Multiracial	0.95%	1.51%
Percentage With Hispanic Ethnicity	6.89%	16.93%
Percentage That Perform Vigorous Exercise at Least Once per Week	28.05%	45.10%

Roffman et al. Scientific Reports, 2017 (under review)

Multi-Parameterized Deep Neural Network



Roffman et al. Scientific Reports, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction

- Sensitivity (true positive rate, or probability of detection) measures the proportion of positives that are correctly identified as positive, = TP/P
- Specificity (true negative rate) measures the proportion of negatives that are correctly identified as negative, = TN/N
- Precision or positive predictive value (PPV), measures how precise is the prediction, = TP/(TP+FP)
- Since the data under-samples prostate cancer, a Bayesian formula is used to calculate the PPV:

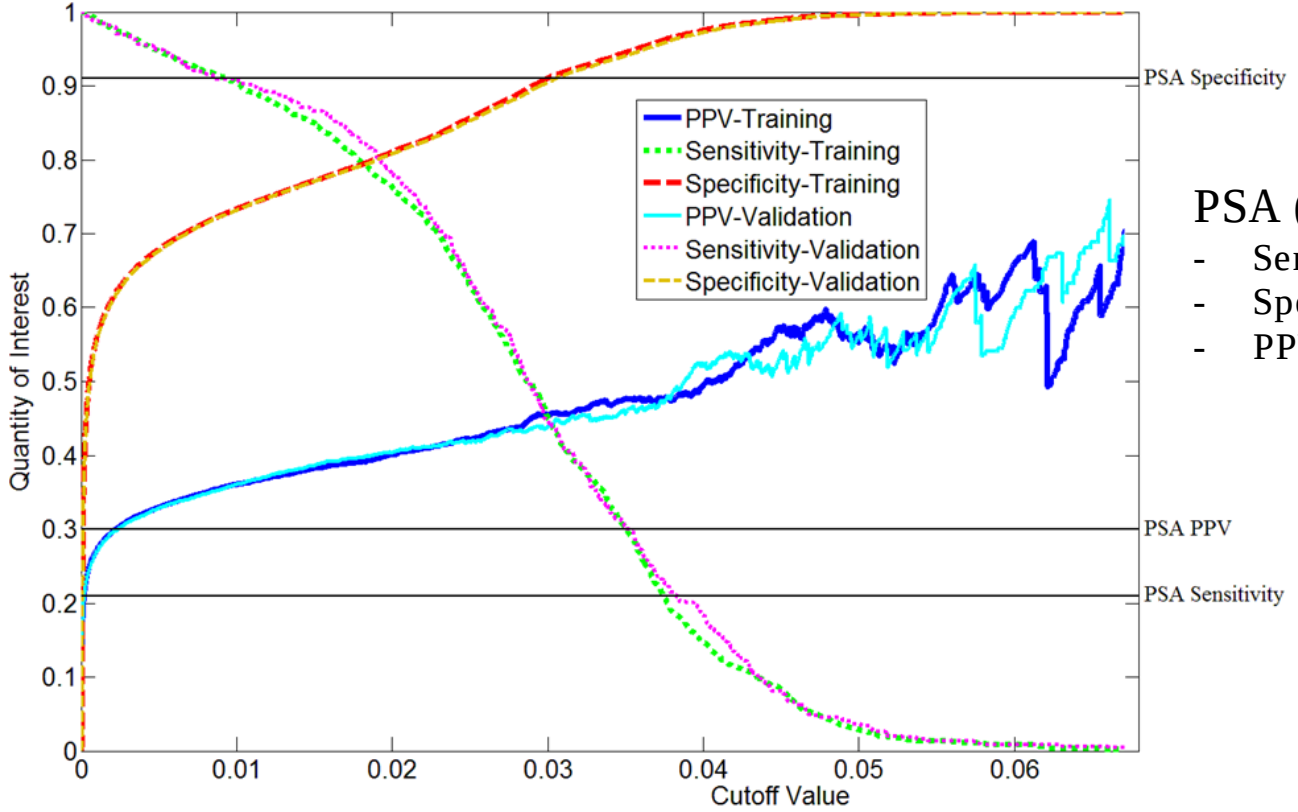
$$PPV = \frac{\text{Sensitivity} * \text{Prevalence}}{(\text{Sensitivity} * \text{Prevalence} + (1 - \text{Specificity}) * (1 - \text{Prevalence}))}$$

Roffman et al. Scientific Reports, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction

DNN training
- Sensitivity: 45%
- Specificity: 91%
- PPV: 46%

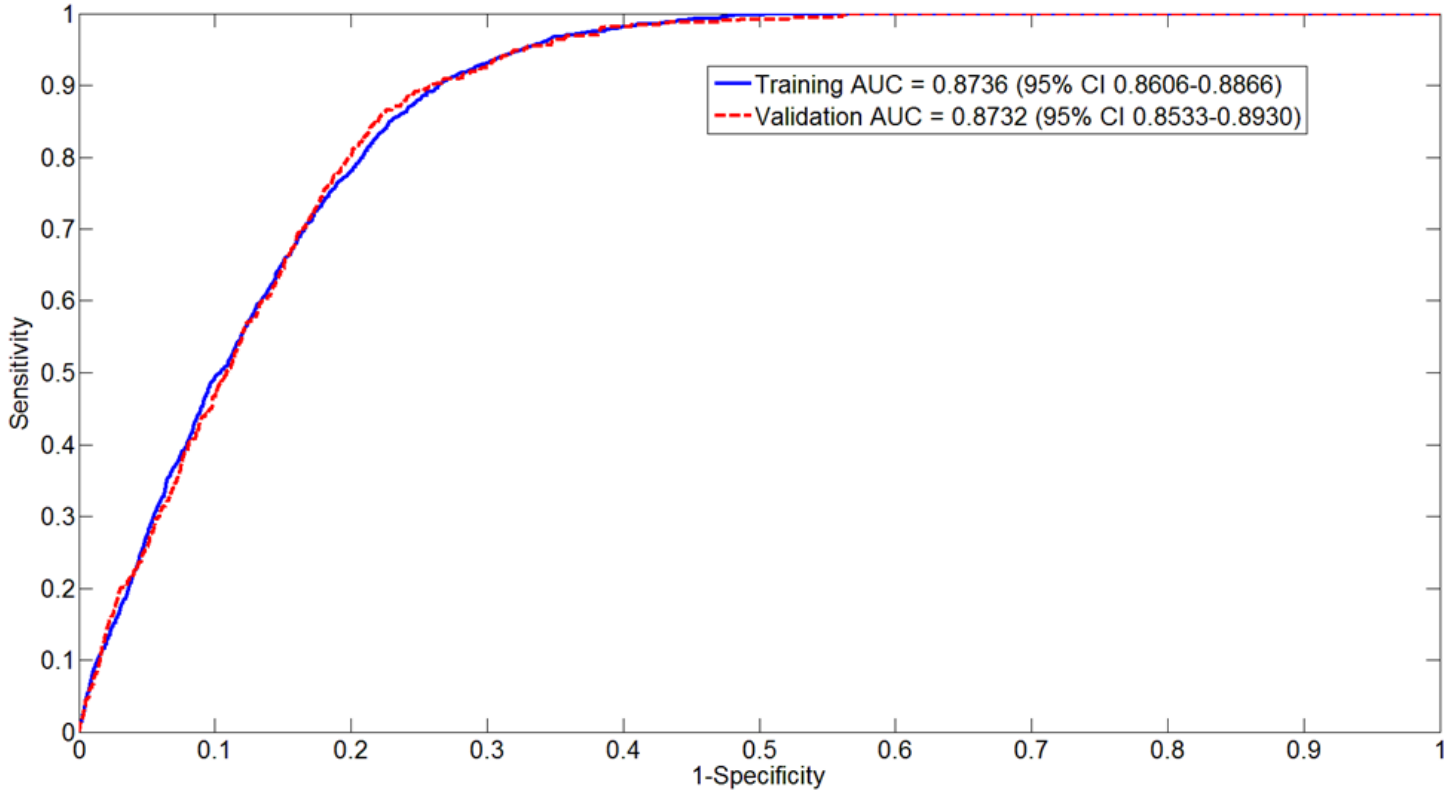
DNN validation
- Sensitivity: 45%
- Specificity: 91%
- PPV: 44%



PSA (ACS)
- Sensitivity: 21%
- Specificity: 91%
- PPV: 30%

Roffman et al. Scientific Reports, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction



Roffman et al. Scientific Reports, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction

Tests	Requirements	Sensitivity	Specificity	AUC
PSA[22,25]	Blood work	95%*	17.2%-19.2%*	0.53-0.549
PHI[25]	Blood work	95%*	36%*	0.815
4- kallikrein score[26,27]	Blood work, prior biopsy, DRE	N/A	N/A	0.82
SelectMDx[23]	Blood work, DRE, urine sample, biomarkers	N/A	N/A	0.86
Clinical Baseline Model[23,30]	Blood work, family history, DRE, prior biopsy	N/A	N/A	0.87
mpMRI[34,35,36]	MRI scan	58%-96% (optimal 95%)	23%-87% (optimal 84%)	N/A
Stockholm-3[33]	Blood work, protein biomarkers, genetic markers, DRE, family history, prior biopsy	N/A	N/A	0.78
22-phage-peptide detector[40]	Serum and unique equipment to conduct the test	81.6%	88.2%	0.93
Radiomics: 5 Haralick texture[38,39,41]	Plethora of imaging data	86%	88%	0.54-0.66
Prostataclass ANN[31,32]	Blood work, DRE, prostate volume measurement	95%	22%-41% (dependent on the PSA value)	0.84
Our ANN	Health informatics commonly available in electronic medical records	95.08%	67.35%	0.8756

- No blood work
- No biopsy
- No imaging
- No genomic data
- No DRE

- Non-invasive
- Cost-effective
- Easy-to-implement

Roffman et al. Scientific Reports, 2017 (under review)

Conclusions

- Big data in radiation oncology is a gold mine waiting to be exploited
- Open data access is the bottleneck to big data applications
- It is crucial to identify which machine learning algorithm is best suited for your specific problem
- It is possible to predict prostate cancer risk for individual with deep neural network based solely on personal health informatics
- There are endless opportunities in machine learning with big health data

You Are Your Data, Your Data is You

In A Digital World

Acknowledgement

NIH/NIBIB: 1R01EB022589-01

Zhe Chen, Ph.D.
James Duncan, Ph.D. (Radiology)
Kenneth Roberts, M.D.
James Yu, M.D.
Yawei Zhang, Ph.D. (Public Health)
Steven Ma, Ph.D. (Biostatistics)
David Roffman, Ph.D.
Liz Guo, MPH
Issa Ali, B.S.
Ying Liang, Ph.D.
Wazir Muhammad, Ph.D.
Gregory Hart, Ph.D.

Thank You!

BIG DATA IN RADIATION ONCOLOGY

Important Dates

Detailed Outline: February 2017
Draft Chapter: **July 1, 2017**

Edited by
Jun Deng, PhD
Associate Professor of Therapeutic Radiology
Yale University School of Medicine
15 York Street, LL508-Smilow
New Haven CT, 06510
+1-860-967-0072 | jun.deng@yale.edu

Lei Xing, PhD
Professor and Director of Medical Physics Division
Department of Radiation Oncology
Stanford University School of Medicine
875 Blake Wilbur Drive, Room G233, Stanford, CA 94305-5847
+1-650-498-7896 | lei@stanford.edu

Taylor & Francis Books, Inc.
Taylor & Francis Group plc