Deep Learning with Big Health Data for Early Cancer Detection and Prevention

Jun Deng, PhD

Professor

Department of Therapeutic Radiology

Yale University School of Medicine

November 4, 2017, Ohio River Valley Chapter Fall Symposium, Indianapolis, IN

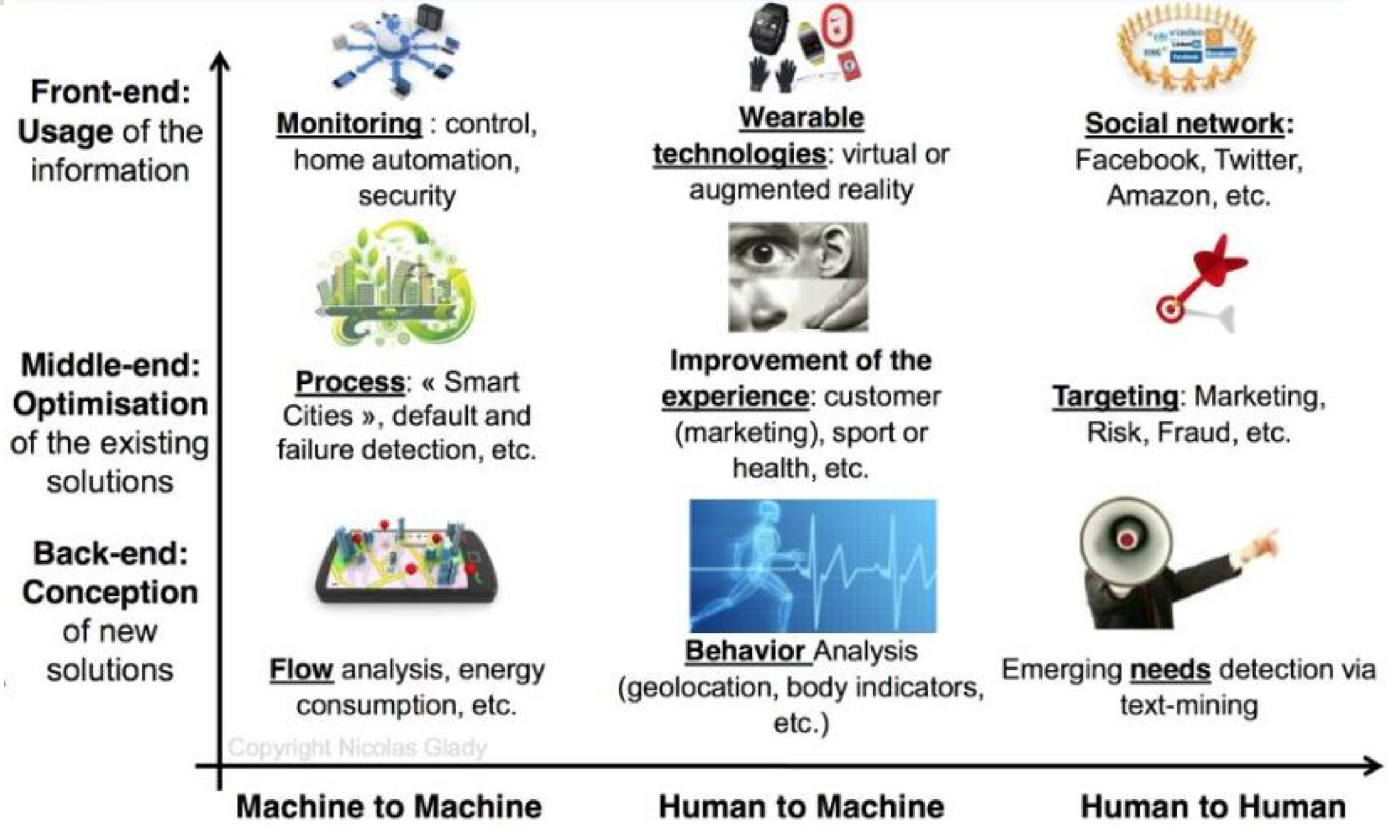
Contents

- Big data basics
- Machine learning 101
- Big data applications in radiation oncology
- Cancer risk prediction via deep learning
- Conclusions
- Future work and 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

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 women 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

More Real World Big Data Applications

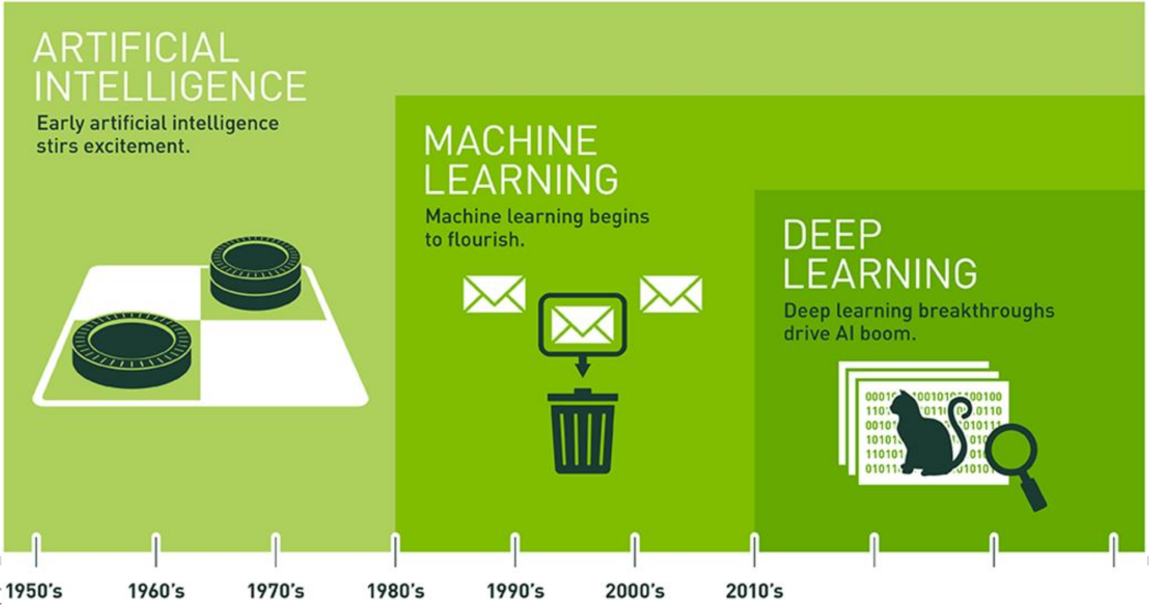
- UPS uses GPS and real-time sensors info to achieve more efficient delivery
- Google forecasts epidemic breakout based on real-time search inquiries
- Amazon recommends books and gift ideas based on your previous choices
- Medtronic predicts hypoglycemic episodes in diabetic patients nearly three hours before its onset, preventing devastating seizures
- Johnson & Johnson analyzes scientific papers to find new connections for drug development
- IBM Watson combs through electronic health records and journal articles from NIH to suggest the best treatment strategy for a cancer patient

Big Data Characteristics

- Four V's: <u>**V**</u>olume, <u>**V**</u>ariety, <u>**V**</u>elocity, and <u>**V**</u>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 <u>**correlation**</u> 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

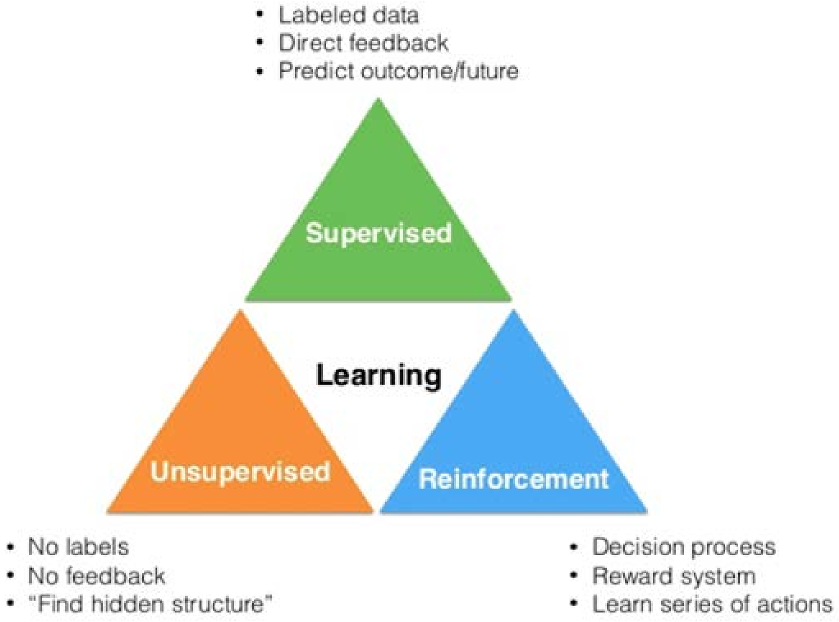


Machine Learning Algorithms

- 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)

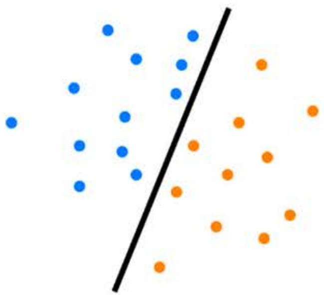
Machine Learning Algorithms

- 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)



Differences and Similarities

Supervised	Unsupervised	Reinforcement
Learning known patterns	Learning unknown patterns	Generating data Learning patterns



“Reinforcement Learning is the true AI”

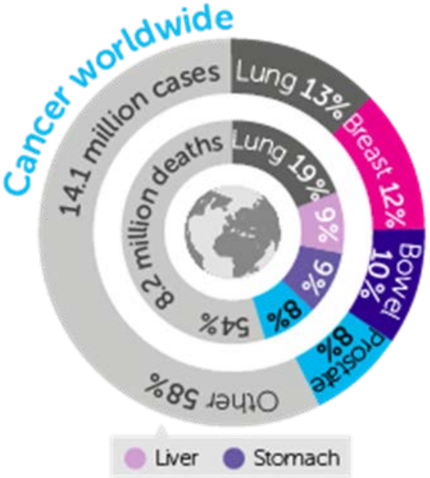
Deep Blue vs Kasparov

- 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

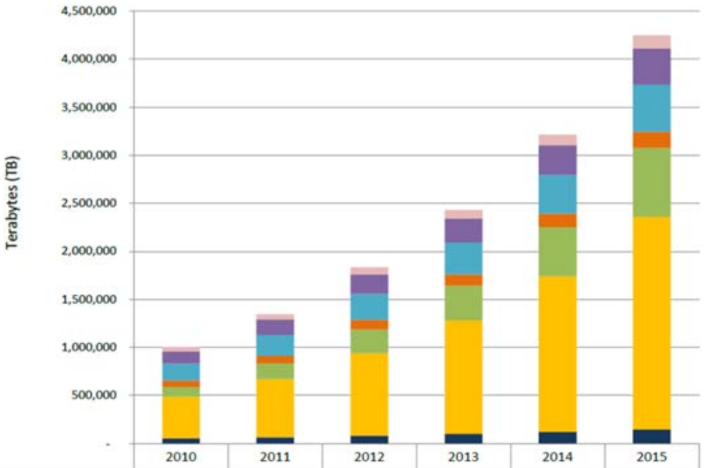
- There are 10^{170} possible positions in Go, too many to try a brute force search
- Google AlphaGo uses deep reinforcement learning to teach the machine to self-learn from its own moves, improve, and make better moves

Cancer Care Big Picture



Oncology
2005-2015
140 M patients
100 K hospitals
0.1-10 GB per patient
14-1400 PB
80% unstructured

Total data, all North American hospitals, by application type, 2010-2015 (TB)



	2010	2011	2012	2013	2014	2015
Research Data	45,007	56,536	72,331	89,876	110,893	137,035
Non-Clinical Imaging	128,307	159,959	202,576	249,808	306,774	375,566
General Unstructured Data/File Services	175,039	216,070	270,544	330,523	402,430	490,478
E-Mail	66,391	80,533	99,176	119,009	142,244	170,060
Electronic Health Records	105,464	163,065	247,852	358,524	508,706	713,673
Clinical Imaging	431,306	603,824	857,499	1,182,290	1,620,810	2,215,525
Administrative Applications	54,518	66,826	82,998	100,388	121,164	146,097

Source: Enterprise Strategy Group, 2011.

Big Data in Radiation Oncology

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

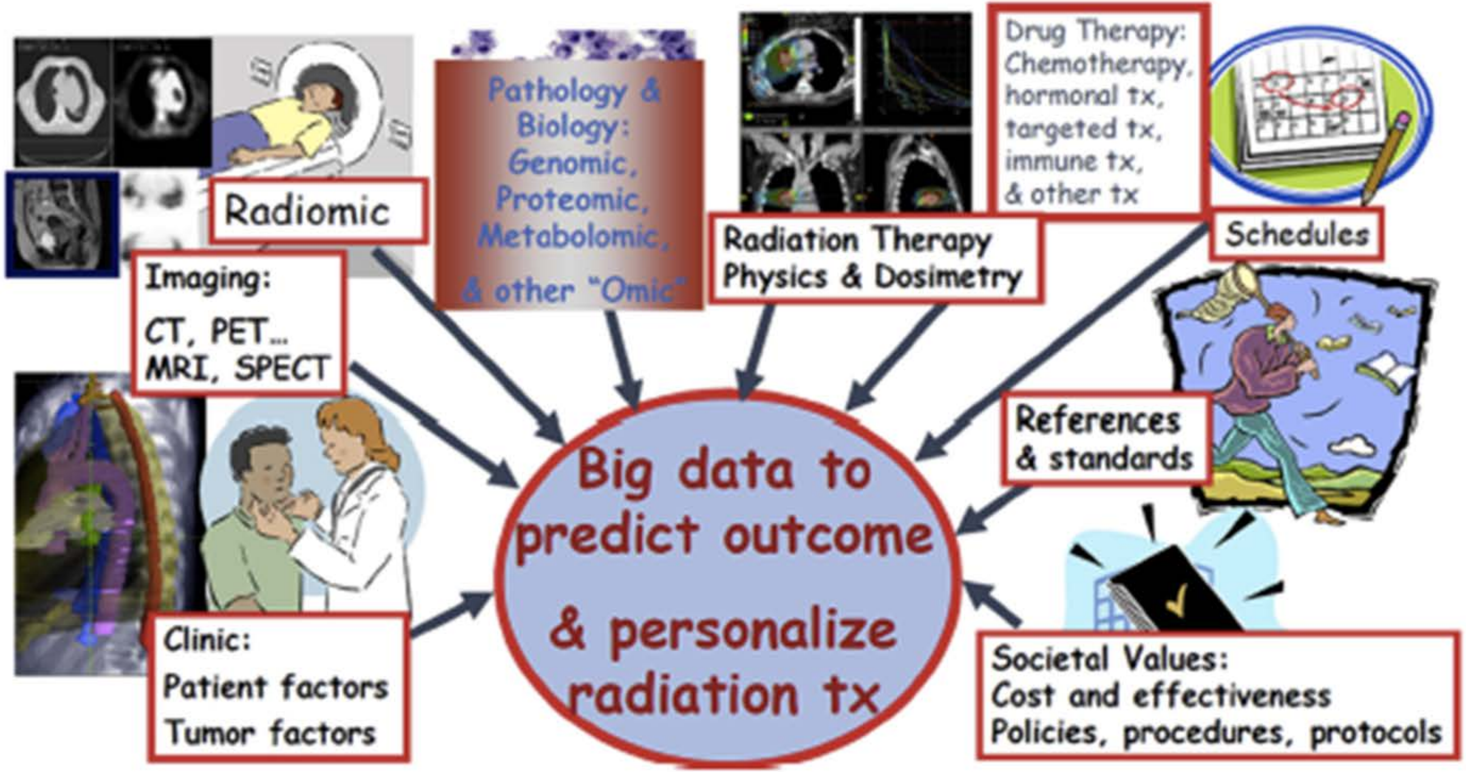
Data type	Data elements	Single patient (average)	Cohort of 1 million patients
Clinical reports	Text	10 MB	10 TB
Laboratory results	Value, units, flag	0.3 MB	0.3 TB
Administrative plus EHR data	Dx, Proc, Rx	2 MB	2 TB
Exome genomic data (variants) (VCF)	Position, type, base(s)	125 MB	125 TB
Imaging data	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

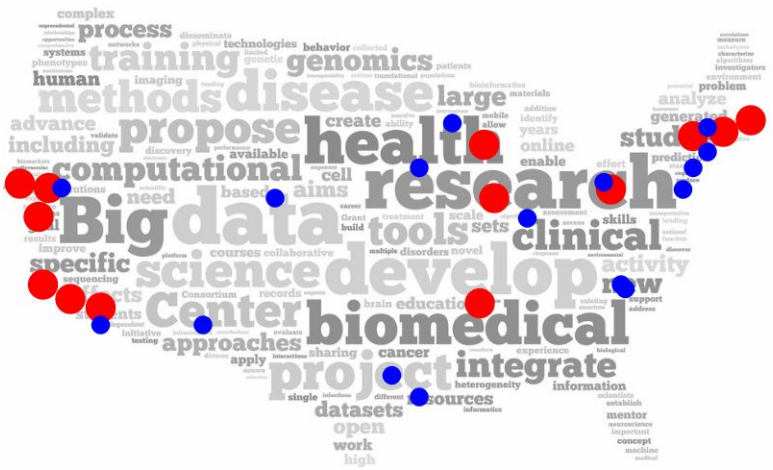
Tap Big Data in Radiation Oncology

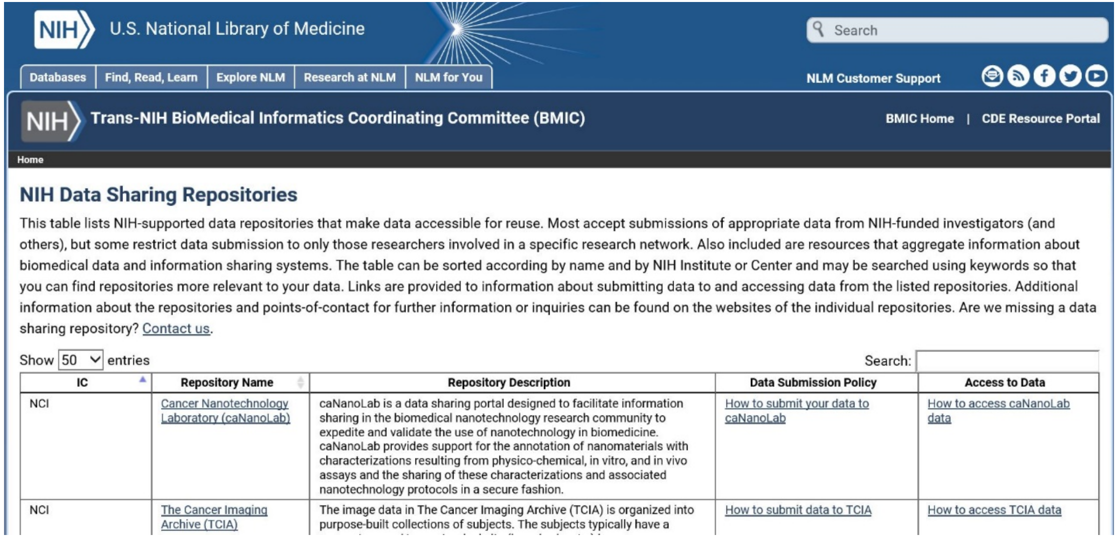


Rosenstein et al, Int J Radiol Oncol Biol Phys 2016

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




NIH U.S. National Library of Medicine

Search

Databases | Find, Read, Learn | Explore NLM | Research at NLM | NLM for You

NLM Customer Support

NIH Trans-NIH BioMedical Informatics Coordinating Committee (BMIC)

BMIC Home | CDE Resource Portal

Home

NIH Data Sharing Repositories

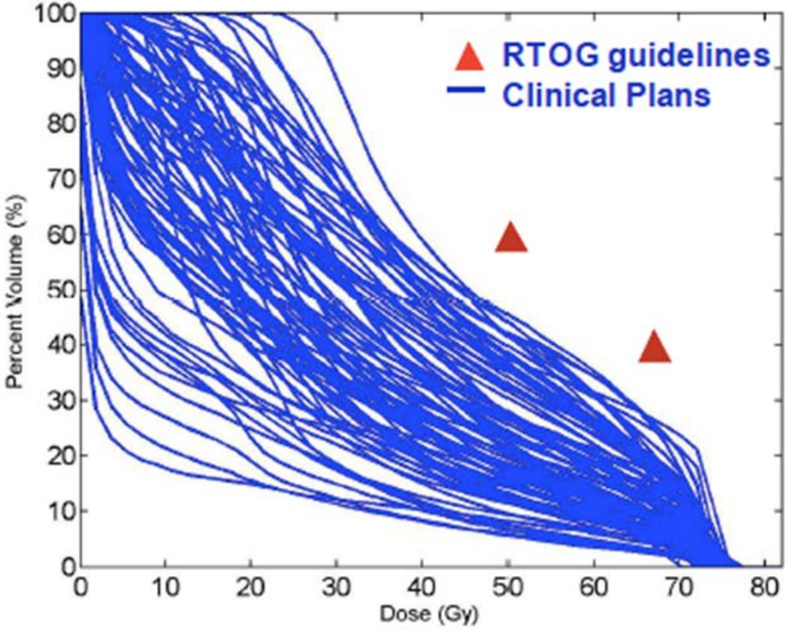
This table lists NIH-supported data repositories that make data accessible for reuse. Most accept submissions of appropriate data from NIH-funded investigators (and others), but some restrict data submission to only those researchers involved in a specific research network. Also included are resources that aggregate information about biomedical data and information sharing systems. The table can be sorted according by name and by NIH Institute or Center and may be searched using keywords so that you can find repositories more relevant to your data. Links are provided to information about submitting data to and accessing data from the listed repositories. Additional information about the repositories and points-of-contact for further information or inquiries can be found on the websites of the individual repositories. Are we missing a data sharing repository? Contact us.

Show 50 entries

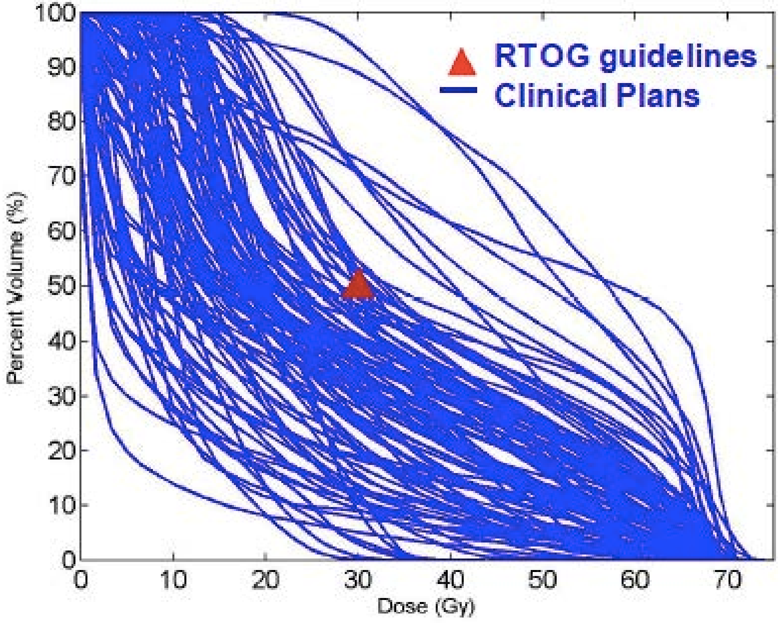
Search:

IC	Repository Name	Repository Description	Data Submission Policy	Access to Data
NCI	Cancer Nanotechnology Laboratory (caNanoLab)	caNanoLab is a data sharing portal designed to facilitate information sharing in the biomedical nanotechnology research community to expedite and validate the use of nanotechnology in biomedicine. caNanoLab provides support for the annotation of nanomaterials with characterizations resulting from physico-chemical, in vitro, and in vivo assays and the sharing of these characterizations and associated nanotechnology protocols in a secure fashion.	How to submit your data to caNanoLab	How to access caNanoLab data
NCI	The Cancer Imaging Archive (TCIA)	The image data in The Cancer Imaging Archive (TCIA) is organized into purpose-built collections of subjects. The subjects typically have a	How to submit data to TCIA	How to access TCIA data

Inter-Plan Variation in IMRT/VMAT



Bladder DVHs/Prostate



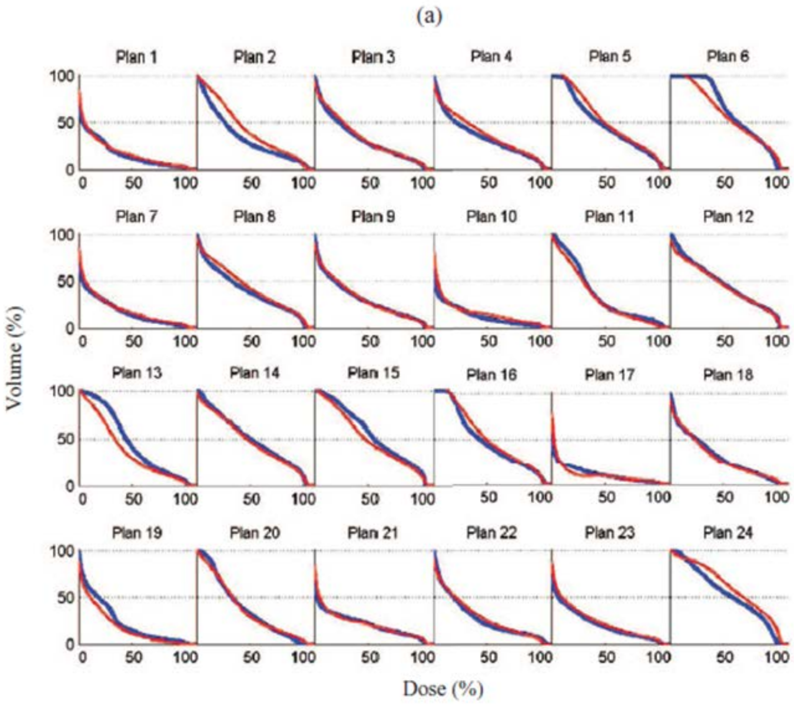
Parotid DVHs/Head & Neck

Courtesy of Huixiao Chen

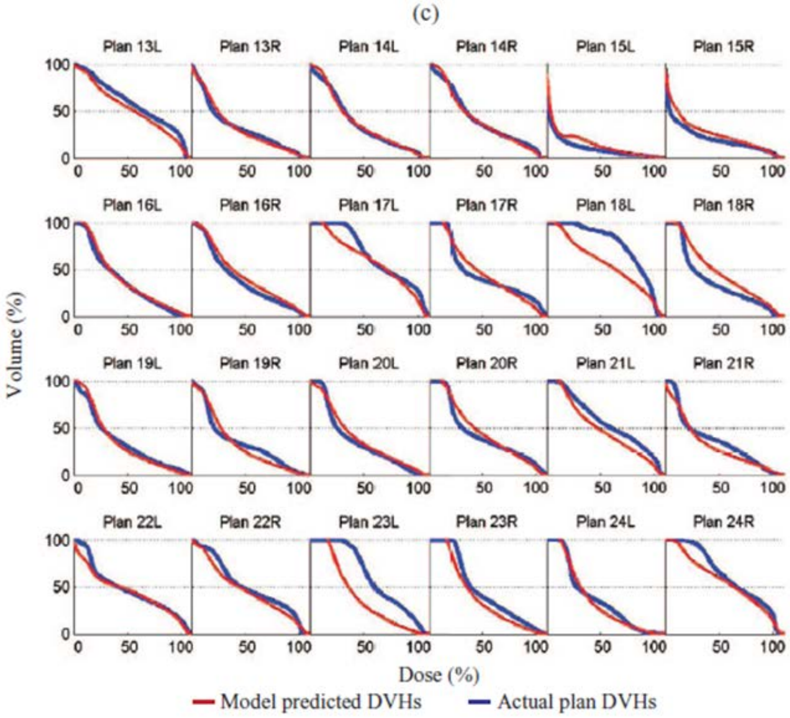
Knowledge-based Treatment Planning

- Based on big data of previous knowledge
- Deep learning for auto-segmentation
- Improved efficiency, reliability, and workflow
- RapidPlan (Varian)
- Pinnacle Auto-Planning (Philips)
- Monaco (Elekta)
- RayStation Automated Planning (RaySearch)

Knowledge-based Treatment Planning



Bladder DVHs/Prostate

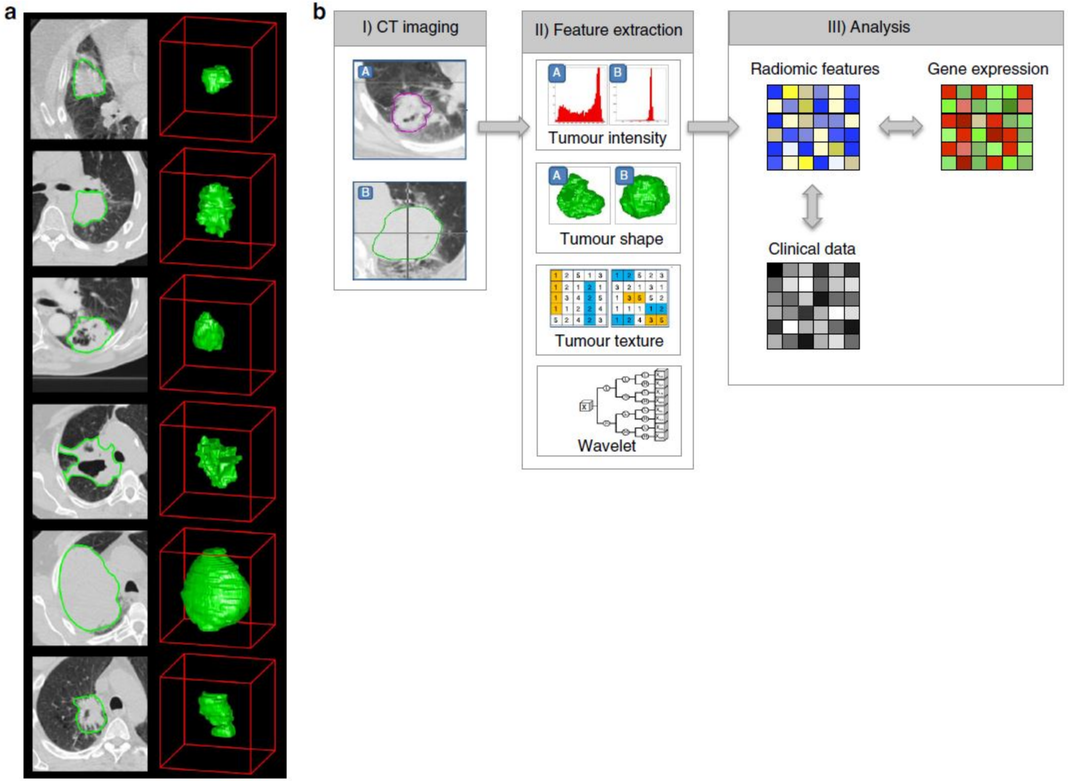


Parotid DVHs/Head & Neck

Radiomics

- Biomarker: a measurable indicator of some biological state or condition
- Biomarker is a *key* element of personalized medicine
  - Prognostic biomarkers: likelihood of disease progression – aggressive vs. indolent
  - Predictive biomarkers: sensitivity to therapy (drugs, radiation)
  - Early response biomarkers: spare patients ineffective treatment; speed up clinical trails
- Radiomics converts imaging data into a high dimensional mineable feature space using automatically extracted data-characterization algorithms
- Hypothesis is that these imaging features capture distinct phenotypic differences of tumors and have prognostic power and clinical significance

Radiomics



Feature Selection Methods	Nnet	DT	BST	BY	BAG	RF	MARS	SVM	NN	GLM	PLSR
RELF	0.51	0.58	0.59	0.72	0.61	0.62	0.53	0.66	0.62	0.79	0.76
FSCR	0.59	0.56	0.6	0.71	0.63	0.63	0.63	0.76	0.57	0.75	0.75
GINI	0.53	0.58	0.58	0.69	0.51	0.56	0.54	0.51	0.65	0.68	0.6
JMI	0.59	0.56	0.52	0.61	0.53	0.59	0.52	0.64	0.65	0.72	0.73
CIFE	0.62	0.55	0.66	0.69	0.68	0.72	0.62	0.77	0.58	0.76	0.78
DISR	0.54	0.53	0.56	0.58	0.51	0.58	0.58	0.53	0.56	0.72	0.6
MIM	0.6	0.53	0.51	0.67	0.55	0.61	0.55	0.72	0.63	0.72	0.75
CMIM	0.74	0.56	0.53	0.58	0.55	0.61	0.6	0.61	0.57	0.72	0.66
ICAP	0.67	0.55	0.55	0.57	0.54	0.6	0.58	0.59	0.65	0.72	0.74
TSCR	0.6	0.61	0.61	0.73	0.53	0.6	0.63	0.71	0.62	0.75	0.75
MRMR	0.53	0.69	0.75	0.7	0.62	0.71	0.64	0.72	0.71	0.58	0.61
MIFS	0.7	0.55	0.55	0.66	0.69	0.68	0.69	0.53	0.63	0.7	0.63
WLCX	0.54	0.5	0.5	0.58	0.56	0.53	0.55	0.55	0.55	0.5	0.7

Classification Methods

Aerts et al. Nat. Commun. 5:4006, doi: 10.1038/ncomms5006 (2014)

Machine Learning for Cancer Prognosis and Prediction

Cancer risk prediction

Publication	Method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Ayer T et al. [19]	ANN	Breast cancer	62,219	Mammographic, demographic	AUC = 0.965	10-fold cross validation	Age, mammography findings
Waddell M et al. [44]	SVM	Multiple myeloma	80	SNPs	71%	Leave-one-out cross validation	snp739514, snp521522, snp994532
Listgarten J et al. [45]	SVM	Breast cancer	174	SNPs	69%	20-fold cross validation	snpCY11B2 (+) 4536 T/C snpCYP1B1 (+) 4328 C/G
Stajadinovic et al. [46]	BN	Colon carcinomatosis	53	Clinical, pathologic	AUC = 0.71	Cross-validation	Primary tumor histology, nodal staging, extent of peritoneal cancer

Cancer survival prediction

Publication	ML method	Cancer type	No of patients	Type of data	Accuracy	Validation method	Important features
Chen Y-C et al. [50]	ANN	Lung cancer	440	Clinical, gene expression	83.5%	Cross validation	Sex, age, T_stage, N_stage LCK and ERBB2 genes
Park K et al. [26]	Graph-based SSL algorithm	Breast cancer	162,500	SEER	71%	5-fold cross validation	Tumor size, age at diagnosis, number of nodes
Chang S-W et al. [32]	SVM	Oral cancer	31	Clinical, genomic	75%	Cross validation	Drink, invasion, p63 gene
Xu X et al. [51]	SVM	Breast cancer	295	Genomic	97%	Leave-one-out cross validation	50-gene signature
Gevaert O et al. [52]	BN	Breast cancer	97	Clinical, microarray	AUC = 0.851	Hold-Out	Age, angioinvasion, grade MMP9, HRASLA and RAB27B genes
Rosado P et al. [53]	SVM	Oral cancer	69	Clinical, molecular	98%	Cross validation	TNM_stage, number of recurrences
Delen D et al. [54]	DT	Breast cancer	200,000	SEER	93%	Cross validation	Age at diagnosis, tumor size, number of nodes, histology
Kim J et al. [36]	SSL Co-training algorithm	Breast cancer	162,500	SEER	76%	5-fold cross validation	Age at diagnosis, tumor size, number of nodes, extension of tumor

Kourou et al, Computational and Structural Biotechnology Journal 13 (2015) 8–17

The Question We Try to Answer

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

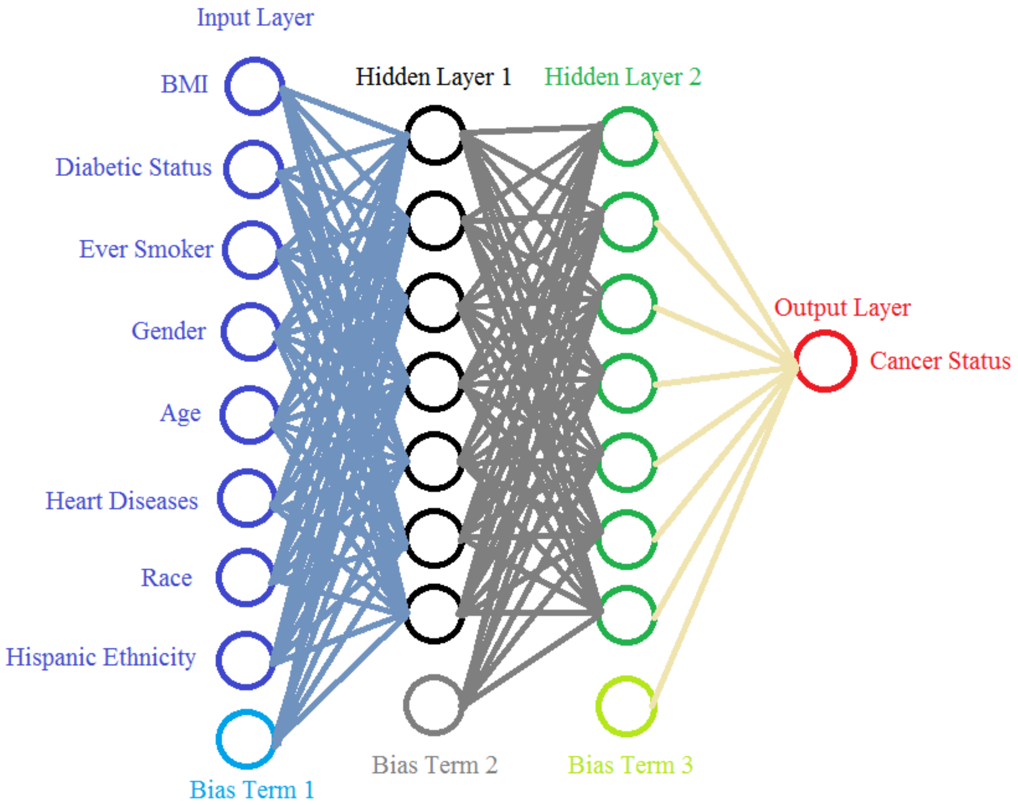
National Health Interview Survey (NHIS)

- Publically available 1997-2015 data
- 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%

Multi-Parameterized Deep Neural Network



Roffman et al. JCO - CCI, 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:

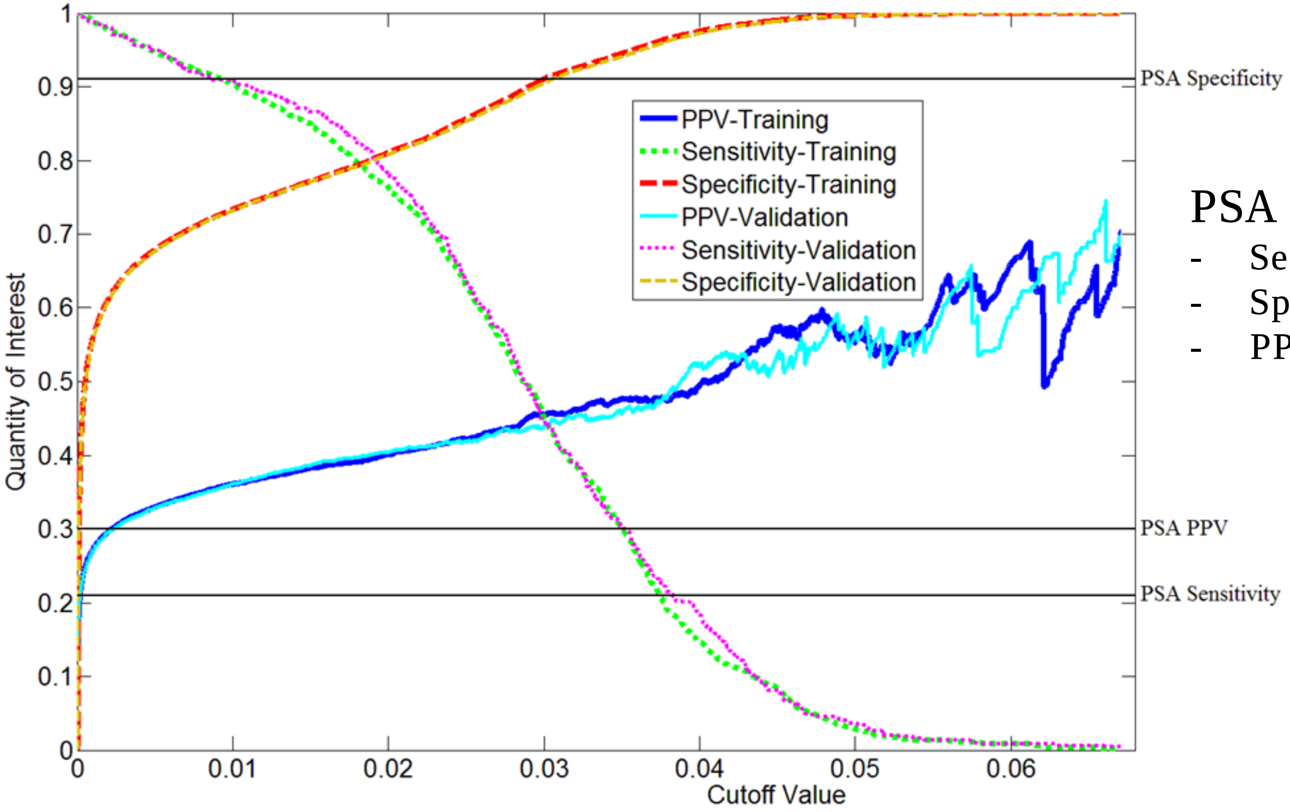
$$PPV = \frac{\text{Sensitivity} * \text{Prevalence}}{(\text{Sensitivity} * \text{Prevalence} + (1 - \text{Specificity}) * (1 - \text{Prevalence}))}$$

Roffman et al. JCO - CCI, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction

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

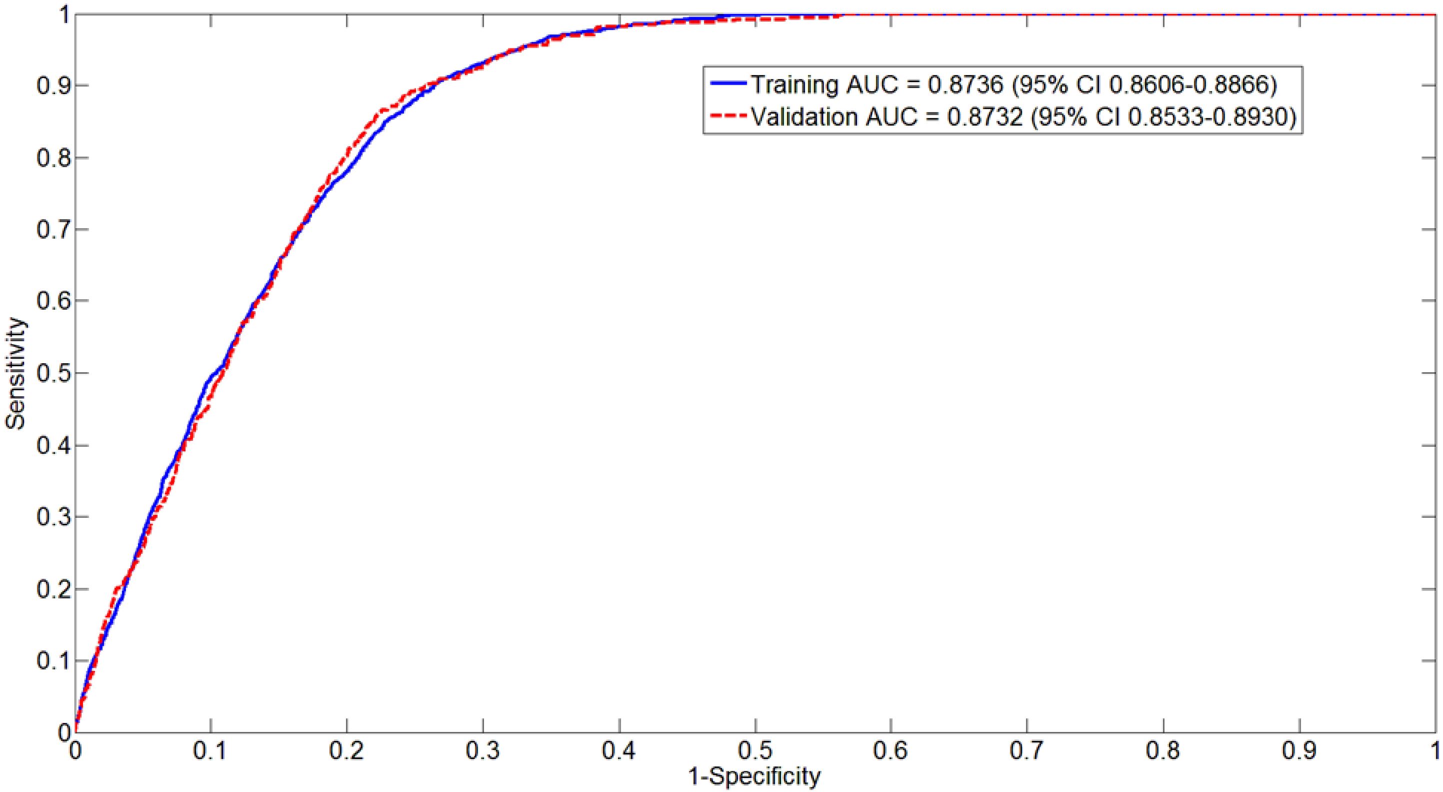
DNN validation
- Sensitivity: 45%
- Specificity: 91%
- PPV: 44%



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

Roffman et al. JCO - CCI, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction



Roffman et al. JCO - CCI, 2017 (under review)

Multi-Parameterized DNN for Prostate Cancer Prediction

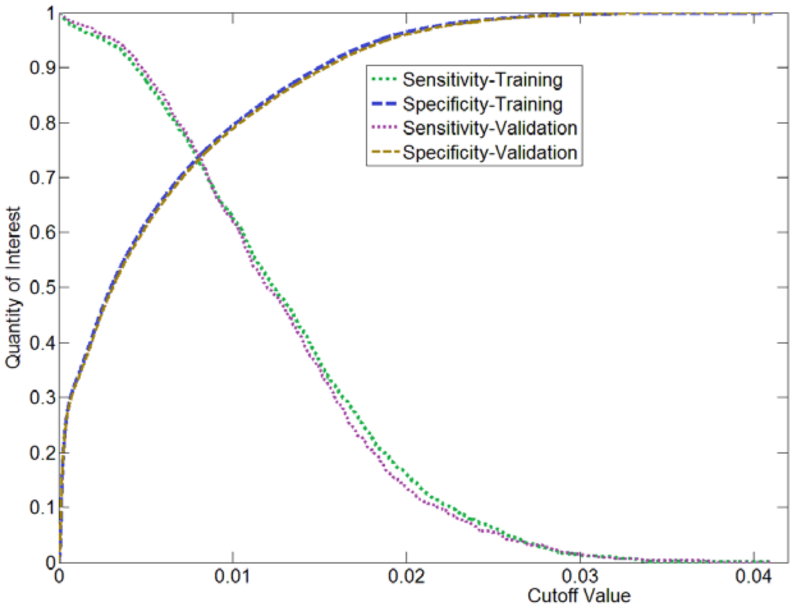
Tests	Requirements	Sensitivity	Specificity	AUC
PSA[22,25]	Blood work	95%*	17.2%-19.2%*	0.53-0.549
PHI[25]	Blood work	95%*	36%*	0.815
4- kallikrein score[26,27]	Blood work, prior biopsy, DRE	N/A	N/A	0.82
SelectMDx[23]	Blood work, DRE, urine sample, biomarkers	N/A	N/A	0.86
Clinical Baseline Model[23,30]	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

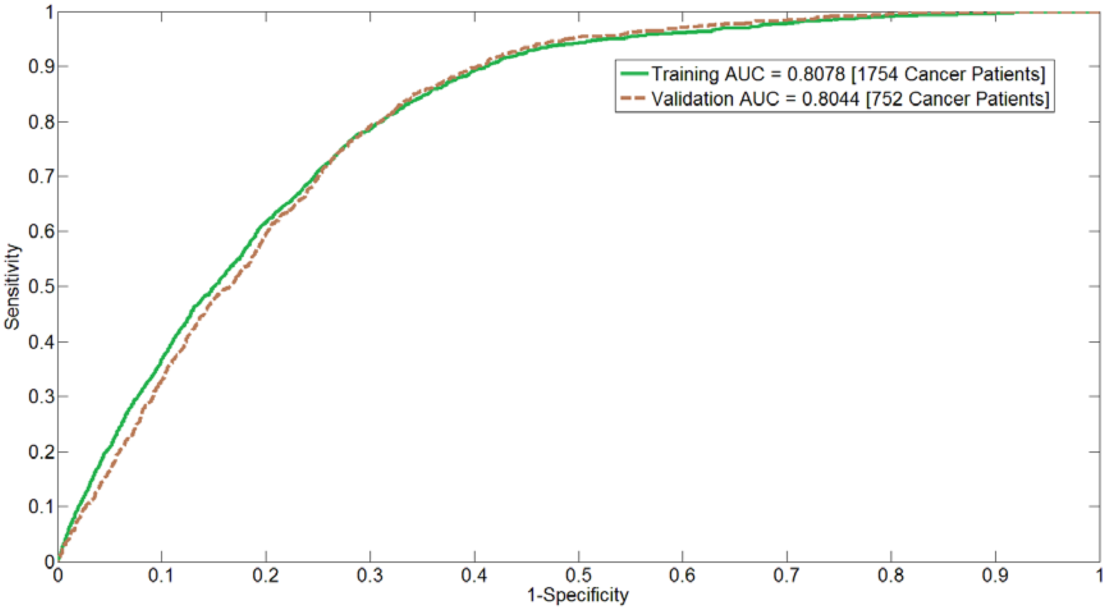
- No blood work
- No biopsy
- No imaging
- No genomic data
- No DRE

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

Roffman et al. JCO - CCI, 2017 (under review)

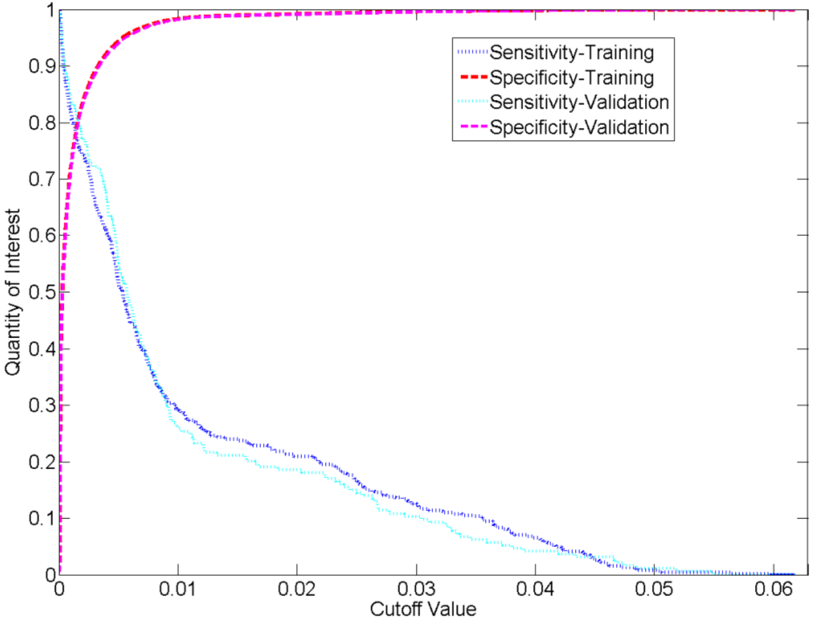
Non-Melanoma Skin Cancer Prediction

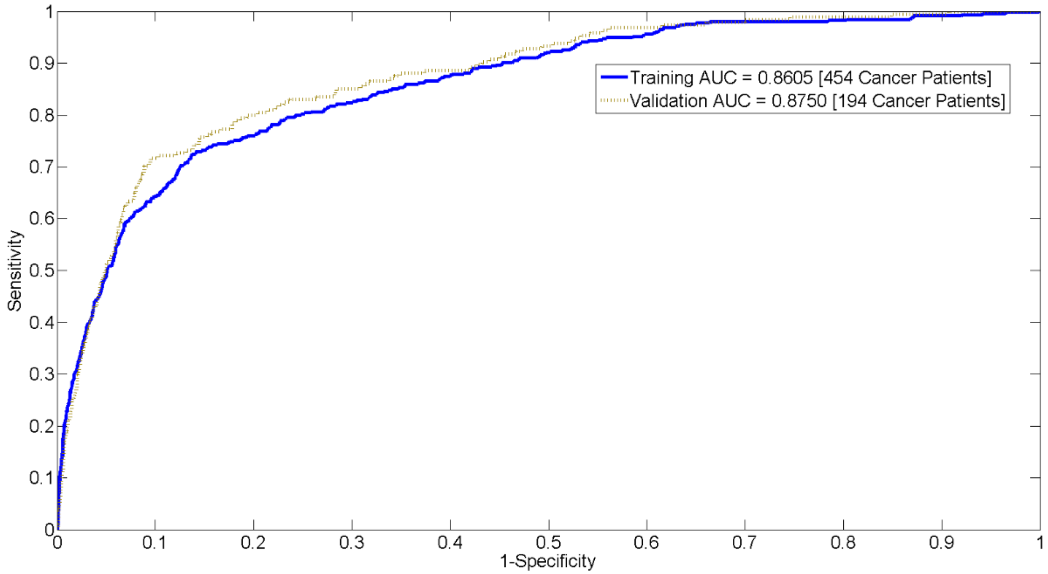




Roffman et al. Frontiers in Oncology, 2017 (to be submitted)

Lung Cancer Prediction

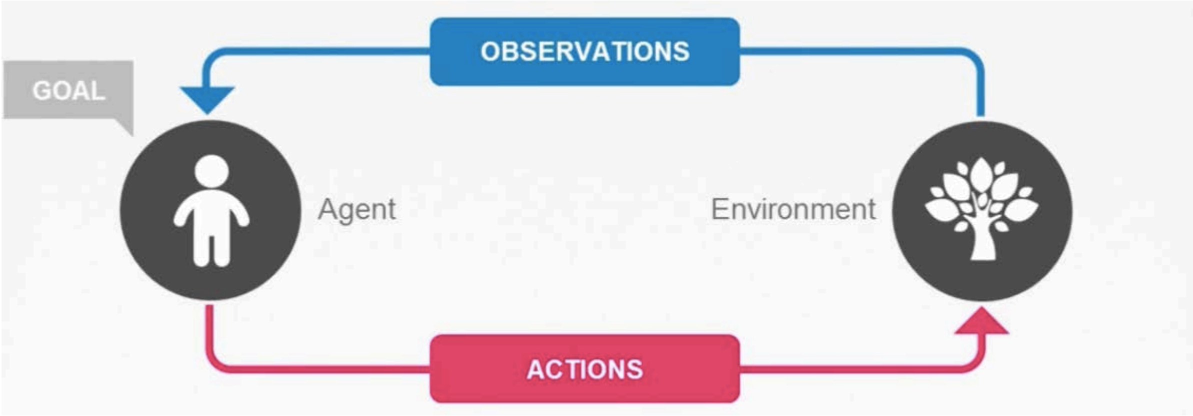




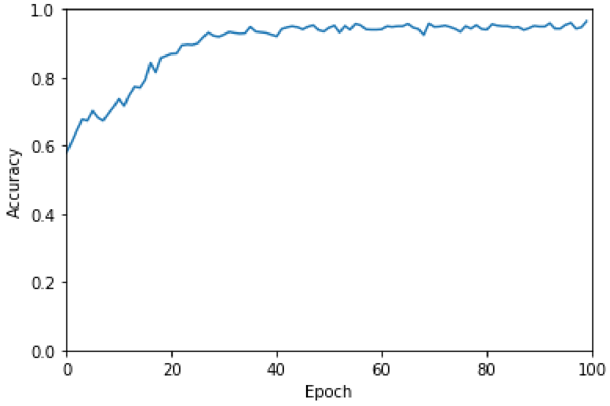
Roffman et al. PLOS ONE, 2017 (to be submitted)

Lung Nodule Detection with Reinforcement Learning

- LUNA (Lung Nodule Analysis) 2016 challenge
  - Publicly available LIDC/IDRI database
  - Annotations based on agreement from minimum 3 out of 4 radiologists
  - Total 888 CT: Nodule = 590 individuals; Non-Nodule = 198 individuals
  - Goal: a large-scale evaluation of automatic nodule detection algorithms
  - https://luna16.grand-challenge.org/



Accuracy



Ali et al. Frontiers in Oncology, 2017 (to be submitted)

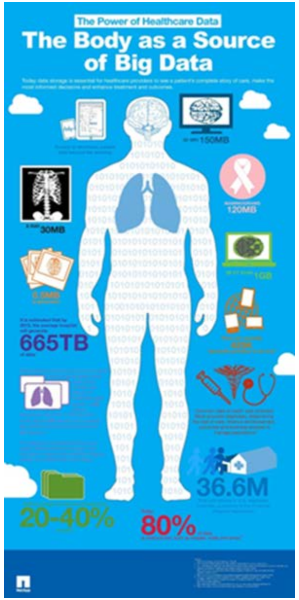
Conclusions

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

Deep Learning of Big Health Data

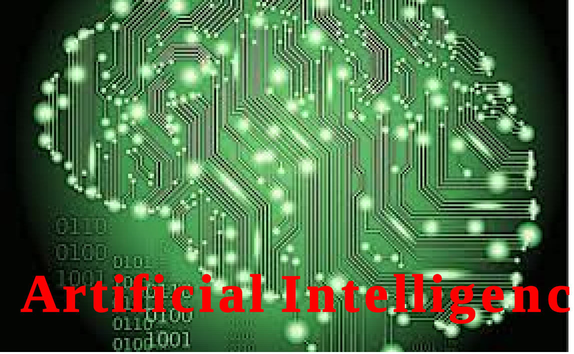
- Health data out-grows the depth and breadth of knowledge any physicians can accumulate in their lifetimes
- Apart from 3% clinical trial data, the remaining 97% is stored in the silo-like EMRs, barely accessible to the physicians as well as the patients who are actually the origin of the data
- Yet, more than 75% of people are currently willing to share their personal health data online for free, with appropriate de-identification
- Meanwhile, AI has shown great promise in tackling big health data to save lives, improve health care and patient outcome, and cut cost

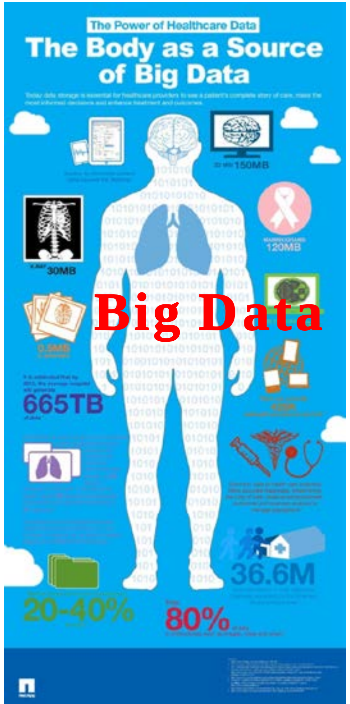
Deep Learning of Big Health Data

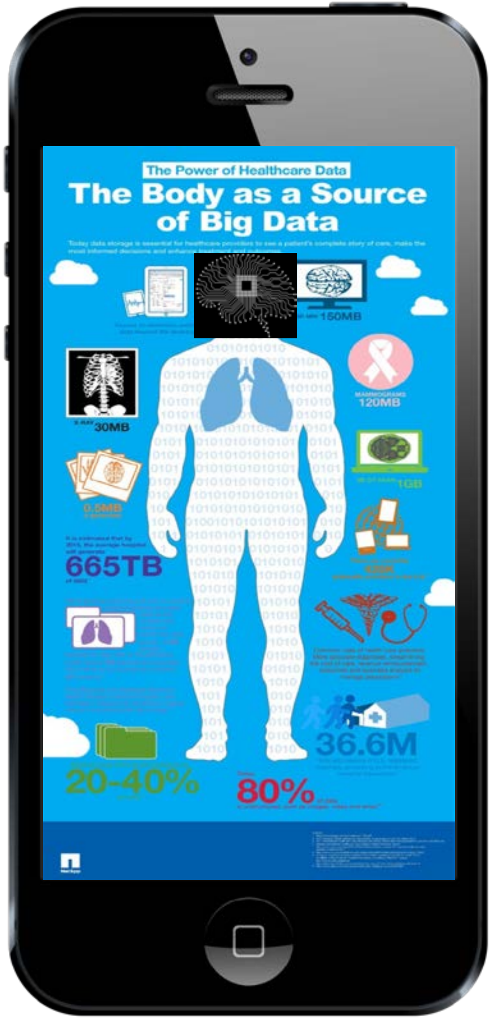


Deep Learning of Big Health Data

- Cultivates a culture of data sharing by strengthening incentives and standards
- Engages patients for effective evidence generation and data sharing for care improvement
- Manages individual cancer risk for the most individualized and effective interventions
- Links the physicians with the patients for shared decision-making







Decentralized Portable Data

You Are Your Data, Your Data is You

In A Digital World

Acknowledgement

NIH/NIBIB: 1R01EB022589-01

Collaborators

Zhe Chen, Ph.D. (Medical Physics)
Kenneth Roberts, M.D. (Rad Onc)
Yawei Zhang, Ph.D. (Public Health)
Michael Leapman, M.D. (Urology)
Christine Ko, M.D. (Dermatology)
Richard Yang, Ph.D. (Computer Science)
Lynn Tanoue, M.D. (Internal Medicine)
Melinda Irwin, Ph.D. (Epidemiology)
James Duncan, Ph.D. (Radiology)
James Yu, M.D. (Rad Onc)
Steven Ma, Ph.D. (Biostatistics)
Roy Decker, M.D. (Rad Onc)
Michael Girardi, M.D. (Dermatology)
Sahand Negahban, Ph.D. (Data Science)
Amy Justice, M.D. (Public Health)
Bonnie Rothberg, M.D. (Med Onc)

Postdocs and Graduate Students

David Roffman, Ph.D.
Issa Ali, B.S.
Wazir Muhammad, Ph.D.
Bradley Nartowt, Ph.D.
Liz Guo, MPH
Ying Liang, 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

Research Topic

Machine Learning with Radiation Oncology Big Data

About this Research Topic

Radiation oncology is uniquely positioned to harness the power of big data as vast amounts of data are generated at an unprecedented pace for individual patients in imaging studies and radiation treatments worldwide. The big data encountered in the radiotherapy clinic may include patient demographics stored in the electronic medical record (EMR) systems, plan settings and dose volumetric information of the tumors and normal tissues generated by treatment planning systems (TPS), anatomical and functional information from diagnostic and therapeutic imaging modalities (e.g., CT, PET, MRI and kVCBCT) stored in picture archiving and communication systems (PACS), as well as the genomics, proteomics and metabolomics information derived from blood and tissue specimens. Yet, the great potential of big data in radiation oncology has not been fully exploited for the benefits of cancer patients due to a variety of technical hurdles and hardware limitations.

With recent development in computer technology, there have been

Topic Editors

Jun Deng
Yale University
New Haven, United States

Issam El Naqa
University of Michigan
Ann Arbor, United States

Lei Xing
Stanford University
Stanford, United States

Q1. The risk factors that may increase a person’s chances of developing cancer include:

a. ionizing radiation to critical organs and tissues

b. environmental conditions such as air quality and chemical absorption

c. lifestyle pattern like smoking, alcohol drinking, and physical activity

d. random mutations during stem cell divisions

e. all of the above

Q1. The risk factors that may increase a person's chances of developing cancer include:

a. ionizing radiation to critical organs and tissues

b. environmental conditions such as air quality and chemical absorption

c. lifestyle pattern like smoking, alcohol drinking, and physical activity

d. random mutations during stem cell divisions

e. all of the above

Answer: e

Reference:

Danaei G, Hoorn SV, Lopez AD et al. *The Lancet*. 2005; 366(9499):1784-1793.
Tomasetti C, Vogelstein B. *Science*. 2015; 347(6217):78-81.

Q2. The main reason(s) that machine learning can be applied in cancer risk prediction is:

a. more and more patient data is accumulated in the clinic routinely and available for mining

b. computer hardware and chip performance has been improved significantly recently

c. there are multiple carcinogenic factors entangled with hidden layers of correlations

d. all of the above

e. none of the above

Q2. The main reason(s) that machine learning can be applied in cancer risk prediction is:

a. more and more patient data is accumulated in the clinic routinely and available for mining

b. computer hardware and chip performance has been improved significantly recently

c. there are multiple carcinogenic factors entangled with hidden layers of correlations

d. all of the above

e. none of the above

Answer: d

Reference:

Bibault J, Giraud P, Burgun A.. *Cancer Letters*. 2016; 382(1): 110-117

Q3. Which V is the biggest problem for extracting big data in radiation oncology?

a. variability

b. velocity

c. volume

d. value

e. none of the above

Q3. Which V is the biggest problem for extracting big data in radiation oncology?

a. variability

b. velocity

c. volume

d. value

e. none of the above

Answer: a

Reference:

Mayo CS, Kessler ML, Eisbruch A. *Advances in Radiation Oncology* (2016) 1, 260-271.

Q4. Which factors are important for enabling incorporation of big data into clinical practice?

a. use of standards

b. database and analytics technologies

c. modifying clinical process to improve availability and curation

d. protecting patient health information

e. all of the above

Q4. Which factors are important for enabling incorporation of big data into clinical practice?

a. use of standards

b. database and analytics technologies

c. modifying clinical process to improve availability and curation

d. protecting patient health information

e. all of the above

Answer: e

Reference:

Mayo CS, Kessler ML, Eisbruch A. *Advances in Radiation Oncology* (2016) 1, 260-271.
McNutt TR, Moore KL, Quon H. *Int J Radiat Oncol Biol Phys*. 2016 Jul 1;95(3):909-15.

Q5. Potentially important source of big data in radiation therapy are:

a. treatment plan and patient data stored in electronic medical record systems

b. insurance claims data

c. RO-ILS

d. b and c above

e. all of the above

Q5. Potentially important source of big data in radiation therapy are:

a. treatment plan and patient data stored in electronic medical record systems

b. insurance claims data

c. RO-ILS

d. b and c above

e. all of the above

Answer: e

Reference:

Potters L, Ford E, Evans S, Pawlicki T, Mutic S. *Int J Radiat Oncol Biol Phys*. 2016 Jul 1;95(3):885-9.