How Much Data Do We Need, and Where Do We Get It?

Joe Deasy, PhD
Department of Medical Physics
Memorial Sloan Kettering Cancer Center

Thanks to

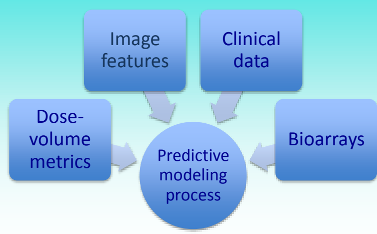
- Harini Veeraraghavan, PhD
- Jung Hun Oh, PhD
- Aditya Apte, PhD
- Maria Thor, PhD
- Mireia Crispin-Ortuzar, PhD
- Andreas Rimner, MD
- Matthew Hellman, MD
- Charles Rudin, MD
- John Humm, PhD
- Margie Hunt, MS
- Amita Dave, PhD
- Neelam Tyagi, PhD
- Nancy Lee, MD
- Heiko Shoeder, MD
- Sang Ho Lee, PhD
- Milan Grkovski, PhD
- And many more…!

Disclosure:

Co-founder of PAIGE.AI, a computational pathology company.

Funding support is gratefully acknowledged:
NIH (R01 and Cancer Center P30 Core Grant),
The Breast Cancer Research Foundation,
Philips, Corp.
Varian Oncology.

Predictive modeling: inputs



Predictive modeling sources and outputs

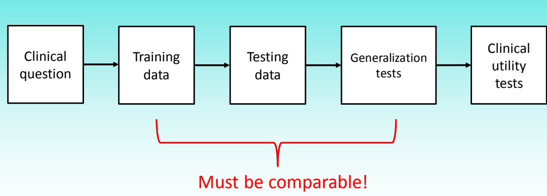
Sources

- Planning dose-volume histograms
- Dose-mapping to identify sensitive regions
- Other planning data: e.g., segmented structures
- Clinical variables
- Radiomics from
  - Diagnostic imaging workup
  - Planning/simulation scan
  - Daily setup images
- Genomic data
  - Tumor actionable mutations
  - RNA, copy number variations
  - Key Germline mutations (BRCA1, BRCA2)
  - Germline genome wide association studies
- Pathology: pathomics

Prediction outputs

- Overall survival
- Local control at a give time point
- Risk of complication (NTCP)
- Likelihood of local control (TCP)
- Genetically stratified risk of complications
- Image segmentation
- Cancer subtype
- Likelihood of response to a given cancer drug
- Likelihood of response to immune therapy
- Risk of developing cancer
- ....

The prediction modeling pipeline



How do you keep from fine tuning the model ***too much*** *in attempting to agree with the input data?*

Cross-validation!

Set-aside cross-validation costs data that could be used for fitting: so why do it?

Rigor: to convince other people (and yourself) that there has been no cheating ('information leakage') that informed the 'hyper parameter' choices.

For small-ish datasets you can use leave-one-out cross validation as the only validation method.

Warning: The entire **supervised** component of the modeling process must be contained within the LOOCV loop, with no prior 'fiddling' with the method. Filing a statistical plan works here as well.

Another way to increase rigor: pre-file your statistical analysis plans (Dekker)

Validation is a higher level of the scientific process than discovery…and it is easier!

Better

1. Internal: good cross validation, **including feature selection**
2. Internal set-aside validation
3. External: same publication
4. External: separate publication/group

Traditional statistical rule of thumb: 10 observations per predictor variable.

Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology. 1996 Dec 1;49(12):1373-9.

So for a modest–sized model (~5 or fewer variables), 50 'events' is probably adequate.

Theoretically: if the model/hypothesis is not known, the bound on the error is not simple:

$$R(h) \leq R_{\text{emp}}(h) + C(|\mathcal{H}|, N, \delta)$$

$$R(h) \leq R_{\text{emp}}(h) + \sqrt{\frac{8d_{vc}(\ln \frac{2m}{d_{vc}} + 1) + 8 \ln \frac{4}{\delta}}{m}}$$

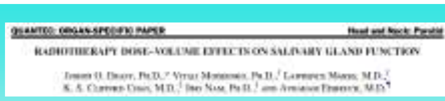
https://mostafa-samir.github.io/ml-theory-pt2/

If the model to be tested is known (Hoeffding's inequality):

This is why validation with a fixed hypothesis is so much easier!

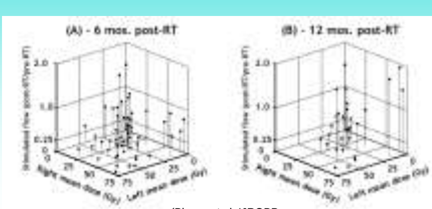
Epsilon is a given bound on the error
m is the size of the dataset
h is a given model/hypothesis

https://mostafa-samir.github.io/ml-theory-pt2/


QUANTEC: ORGAN-SPECIFIC PAPER

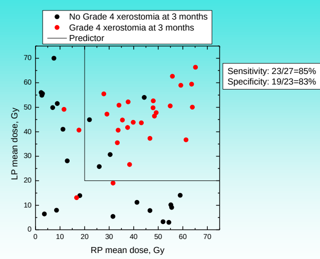
Head and Neck: Parotid

RADIOTHERAPY DOSE-VOLUME EFFECTS ON SALIVARY GLAND FUNCTION



(Blanco et al. IJROBP 2005)

Test of QUANTEC "20/20" recommendation



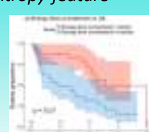
(Unpublished data, Vitali Moiseenko, John Wu, et al., BCCA)

Radiomics: "*Immunotherapy benefit for lung cancer patients is associated with tumor heterogeneity determined from computed tomography radiomic entropy feature"*

Approach:

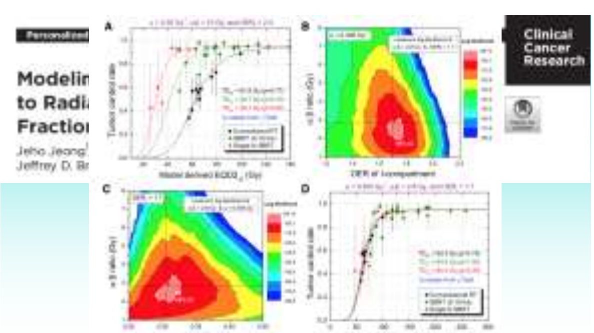
- measures of heterogeneity predict late (> 1 yr.) immunotherapy response.
- 62 NSLC patients treated with pembrolizumab.
- Four radiomic measures of heterogeneity were extracted from longitudinal CT scans, including entropy of values over small patches.
- High entropy implies neighboring voxels are relatively dissimilar in intensity.

Key result:



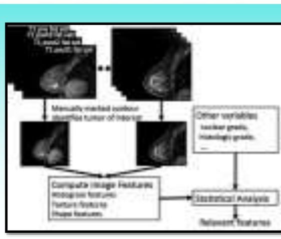
- High entropy at first treatment scan predicts durable response.
- The change in entropy from baseline is more important than baseline entropy.

H. Veeraraghavan[1], M. Hellmann[2], H. Rizvi[2], J. Jiang[1], D. Halpenny[3], A. Snyder[2], ..., J. Deasy[1], MSKCC Depts. of [1] Medical Physics, [2] Medicine, and [3] Radiology (in review)

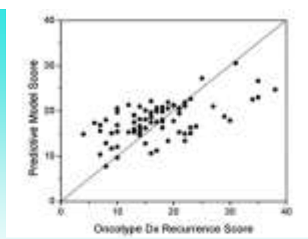


Breast cancer subtype intertumor heterogeneity: MRI-based features predict results of a genomic assay.

Sutton EJ, Oh JH, Dashevsky BZ, Veeraraghavan H, Apte AP, Thakur SB, Deasy JO, Morris EA



N=95



$$Y\ (\mathrm{ODxRS}) = -5.95 \times \mathrm{kurtosis}[\mathrm{Postcontrast}\ 1] + 6.29 \times \mathrm{kurtosis}[\mathrm{Postcontrast}\ 3] + 3.90 \times \mathrm{nuclear\ grade} + 4.12$$

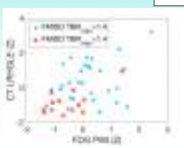
R^2=0.23

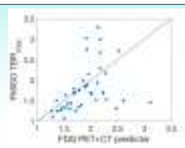
Do we know the best machine learning tools for radiomics

- "No," but this is probably less of an effect compared to
  - variability between imaging systems/calibration practices/protocols
  - Data starvation
- Low dimensional modeling should always be tried
- If dataset is particular rich, higher dimensional data analysis may be justified, with careful control of the risk of overfitting

Combined PET and CT radiomics features predict maximum FMISO uptake in head and neck cancer (Crispin-Ortuzar et al., Radiother. Oncol., 2018)

- FDG PET + contrast-enhanced CT to predict maximum FMISO TBR
- 79 training, 42 hold-out validation
- LASSO + 10x10-fold CV
- Selected predictors:
  - *P90 FDG SUV*
  - *Long run high grey level emphasis in low-FDG subregion*
- Validation AUC = 0.83





Radiomics analysis of pulmonary nodules in low-dose CT for early detection of lung cancer

Wookjin Choi, Jung Hun Oh, and Sadegh Riyahi

Chia-Ju Liu

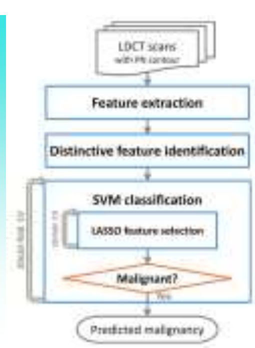
Feng Jiang

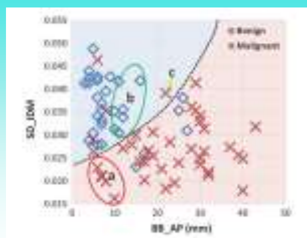
Wengen Chen and Charles White

Andreas Rimner

James G. Mechalakos, Joseph O. Deasy, and Wei Lu*

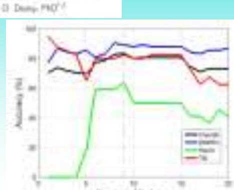
(Medical Physics, 2018) N=72





Breast Cancer Molecular Subtype Classifier That Incorporates MRI Features

J. MAGN. RESON. IMAGING 2016;44:122-129.

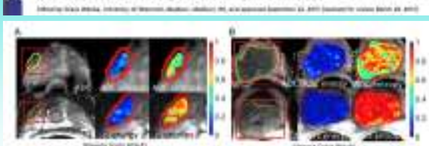


N=178
LOOCV

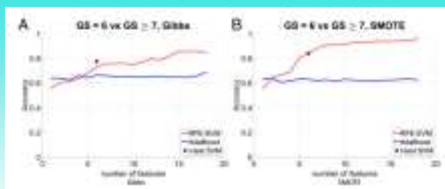
Prediction accuracy for IDC subtype was 81.4%

PNAS

Automatic classification of prostate cancer Gleason scores from multiparametric magnetic resonance images



N=217



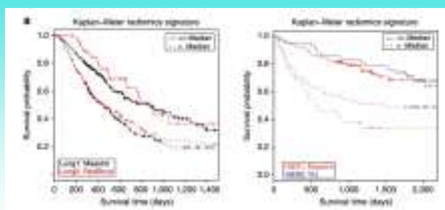
A key was augmenting the learning dataset with simulated data.

ARTICLE

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

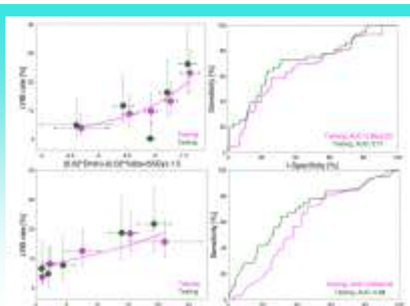
N=474, training
324, testing (NSCLC)
231, testing (H&N)
7 cohorts

Note the emphasis on reproducible features and dimensional reduction!



Pooled cohort analysis demonstrates the importance of rectal sparing in preventing late rectal bleeding

- 989 patients, 5 institutions
- treated with 3DCRT or IMRT to
- 70-86.4Gy@1.8-2.0Gy/fraction



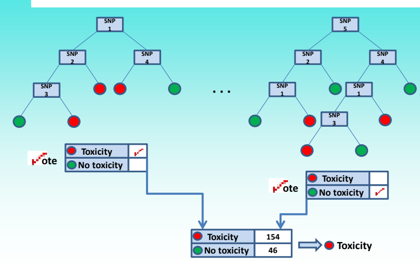

SCIENTIFIC REPORTS

OPEN Computational methods using genome-wide association studies to predict radiotherapy complications and to identify correlative molecular processes

Training dataset
- 243 samples
- 49 events
- 749 SNPs (p< 0.001; Chi-square test)

- Validation dataset
- 122 samples
- 25 events

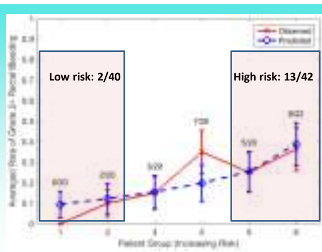
Random Forest



Dataset for RB

- 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 CV or bootstrapping with 100 iterations
- Additive model
  - Coded as the number of rare alleles

Results for RB using validation data



"Deep learning enables fully automatic tumor segmentation and synthesis of MRI from CT images"

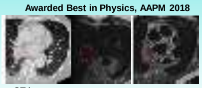
Successfully segmented **1300** NSCLC tumors from three datasets irrespective of tumor size, location, and malignancy.

Our method learns to synthesize MRI from CT images by training with highly limited number of MRI (n=9) with unrelated CT scans (n=300)

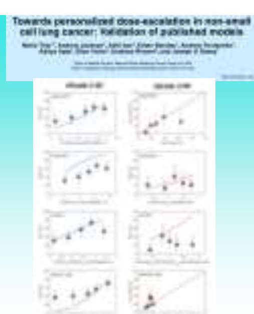
Our method has auto-segmented largest number of lung tumors published

Red – Expert, Blue - Algorithm

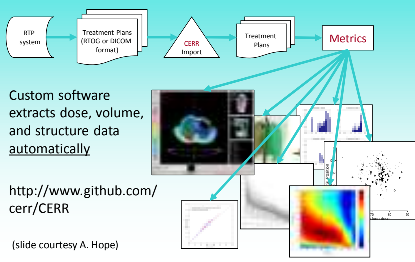
Awarded Best in Physics, AAPM 2018



Slide courtesy: Harini Veeraraghavan, Jue Jiang



The Computational Environment for Radiotherapy Research (CERR)



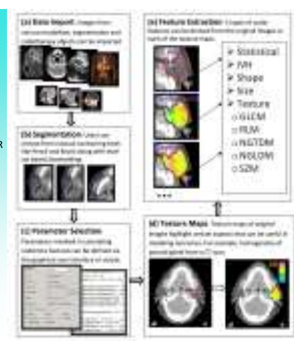
- Custom software extracts dose, volume, and structure data automatically
- http://www.github.com/cerr/CERR

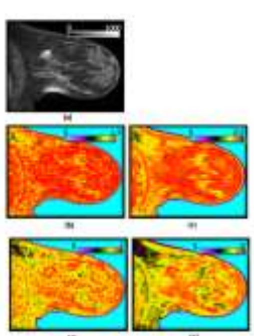
(slide courtesy A. Hope)

40

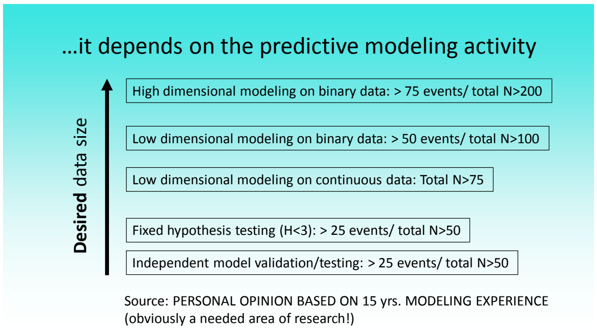
Radiomics Image Quantifier (RIQ) Toolbox in CERR, Aditya Apte et al.

www.github.com/cerr/CERR





Different ways to compute Haralick entropy lead to different feature maps


…it depends on the predictive modeling activity
Desired data size
High dimensional modeling on binary data: > 75 events/ total N>200
Low dimensional modeling on binary data: > 50 events/ total N>100
Low dimensional modeling on continuous data: Total N>75
Fixed hypothesis testing (H<3): > 25 events/ total N>50
Independent model validation/testing: > 25 events/ total N>50
Source: PERSONAL OPINION BASED ON 15 yrs. MODELING EXPERIENCE
(obviously a needed area of research!)