Clifton (Dave) Fuller, MD, PhD
Assistant Professor
Head & Neck Section

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Machine Learning, "Big Data", Informatics, Radiomics, and Clinical Outcomes

The Future of Innovation in Clinical Radiation Oncology?

C.D. Fuller Acknowledgment/Disclosures

2016-17 Funders:

- The Andrew Sabin Family Fellowship Program, through an endowment established by the Andrew Sabin Family Foundation
- A direct gift from the Beach Family of Phoenix, AZ.
- NIH Big Data to Knowledge (BD2K) Program of the National Cancer Institute (NCI) Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Award (1R01CA214825-01)
- National Science Foundation, Division of Mathematical Sciences, Quantitative Approaches to Biomedical Big Data (QuBBD)/Big Data to Knowledge (BD2K) Program (NSF1557559;
- National Cancer Institute Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science (1 R01 CA214825-01)
- National Institute of Dental and Craniofacial Research (NR56/R01 DE025248-01)
- National Cancer Institute Grant MD Anderson Head and Neck Specialized Programs of Research Excellence (SPORE) Development Award (P50CA097007-10)
- Elekta AB/MD Anderson MRI-LinAc Consortium Seed Grant*
- Elekta AB Travel support &Honoraria*

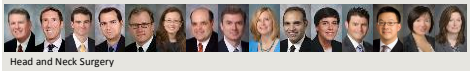
#-Philantropic individuals/agencies/societies
*-Corporate/industry funders
+-Federal or state funding agencies

Radiation Oncology Head and Neck Section

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History

GSBS

MDACC Head and Neck Team



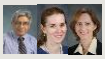
Head and Neck Surgery

Speech Pathology

Thoracic/Head and Neck Medical Oncology

Neuroradiology

Radiation Oncology/Medical Physics

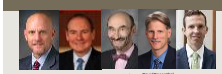


Pathology

Oncologic Dentistry

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Institutional/Departmental Team



Institutional/Departmental/Section/Residency Leadership

Supporting Program/Center Leadership

Research/Grant Team

THE UNIVERSITY OF TEXAS
MD Anderson Cancer Center
Making Cancer History®

Clinic Team

Admin

GSBS

MD Anderson Multi-disciplinary Symptom Working Group

Stephen Lai
MD, PhD
Head and Neck Surgery

Kate Hutcheson
PhD
Speech Pathology

Jinzhong Yang
PhD
Radiation Oncology

bdallah Mohamed
MD, MSc
Radiation Oncology

Jihong Wang
PhD
Radiation Oncology

Dave Fuller
MD, PhD
Radiation Oncology

•National Institute of Dental and Craniofacial Research (NR56/R01 DE025248-01; SY Lai, PI)

•National Cancer Institute Grant MD Anderson Head and Neck Specialized Programs of Research Excellence (SPORE) Development Award (P50CA097007-10, J Myers, PI)

•National Institutes of Health/National Cancer Institute Grant (R03 CA188162-01A1; KA Hutcheson, PI)

2014 Statistical and Applied Mathematical Sciences Institute (SAMSI) Innovation Lab
"Interdisciplinary Approaches to Biomedical Data Science Challenges" Team

Liz Marai
PhD
Computer Science
U. Illinois-Chicago

David Vock
PhD
Biostatistics
U. Minnesota

Guadalupe Canahuate
PhD
Computer Science
U. Iowa

Dave Fuller
MD, PhD
Radiation Oncology
UT MD Anderson

-2016 Joint NIH/NSF Division of Mathematical Sciences Initiative on Quantitative Approaches to Biomedical Big Data (QuBBD) Grant, *"Spatial-nonspatial Multidimensional Adaptive Radiotherapy Treatment"* (NSF 1557679)
- 2017-2020 Early Stage Development of Technologies in Biomedical Computing, Informatics, and Big Data Science Grant, *"SMART-ACT: Spatial Methodologic Approaches for Risk Assessment and Therapeutic Adaptation in Cancer Treatment"* (R01 CA214825-01)

National Science Foundation
WHERE DISCOVERIES BEGIN

Personal Mentor Team

Pre-medical, medical, graduate, residency, thesis committee, and post-doctoral mentors

Current scientific and clinical mentorship team

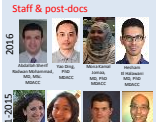
David I. Rosenthal, MD

Charles R. Thomas, Jr., MD

Long-term career mentors (15+ years)

Trainee Team

Staff & post-docs



Resident Physicians

Medical and Graduate Student Trainees

Undergraduate Trainees

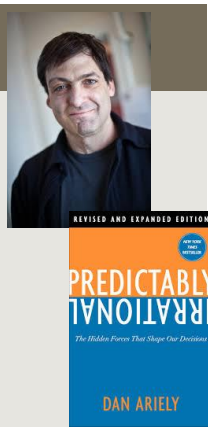
The Home Team

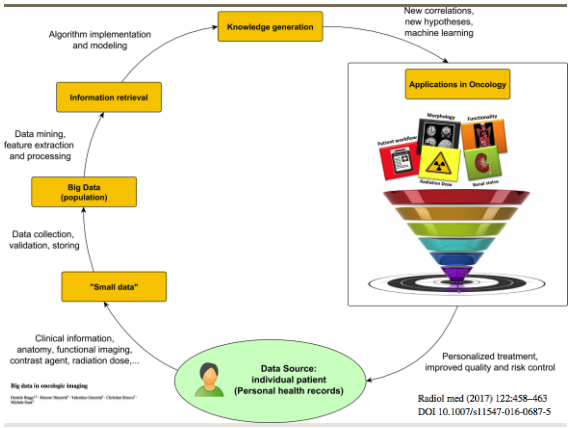
A few definitions…

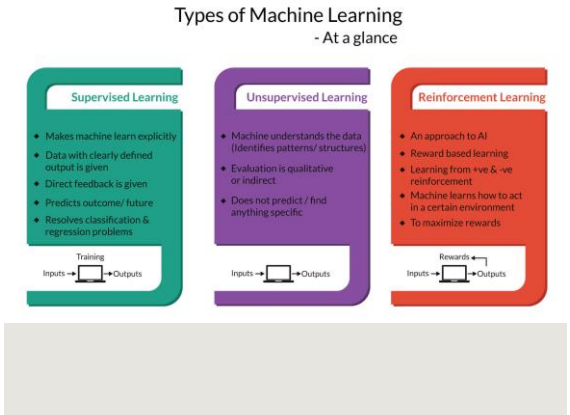
- Informatics [The Discipline]
  - "Informatics is the science of processing data for storage and retrieval; information science as a field."
- "Big Data" [The Input]
  - "*Big Data* is high-volume, high-velocity and/or high-variety informatic assets that demand cost-effective, innovative forms of information processing that enable enhanced insight, decision making, and proc automation."- Gartner
- Machine learning ("statistical learning") [The Methodology]
  - "Machine learning explores the study and construction of algorithms that ca learn from and make predictions on data – such algorithms overcome follov strictly static program instructions by making data-driven predictions or decisions, through building a model from sample inputs."

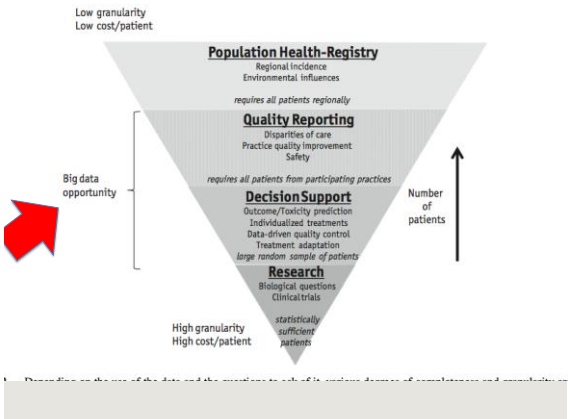
"Big Data"

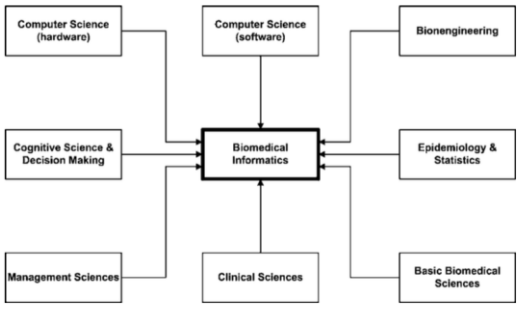
- "Big data is like teenage sex: everyone talks about it, nobody really knows how to do it, everyone thinks everyone else is doing it, so everyone claims they are doing it too…and most don't do it very well."
  – Dan Ariely



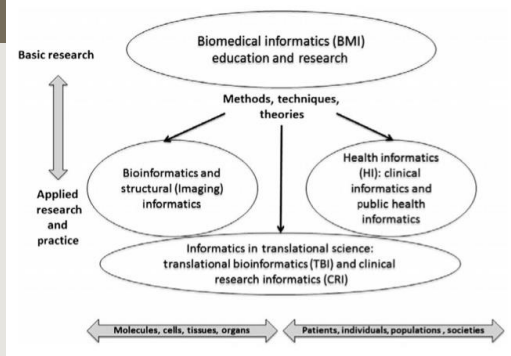

Knowledge generation
New correlations, new hypotheses, machine learning
Algorithm implementation and modeling
Applications in Oncology
Information retrieval
Data mining, feature extraction and processing
Big Data (population)
Data collection, validation, storing
"Small data"
Clinical information, anatomy, functional imaging, contrast agent, radiation dose,...
Data Source: individual patient (Personal health records)
Personalized treatment, improved quality and risk control
Big data in oncologic imaging
Radiol med (2017) 122:458–463
DOI 10.1007/s11547-016-0687-5

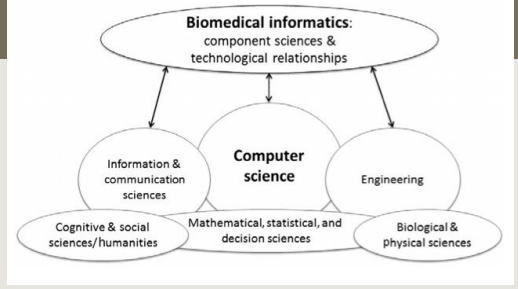

Types of Machine Learning
- At a glance
Supervised Learning
Makes machine learn explicitly
Data with clearly defined output is given
Direct feedback is given
Predicts outcome/ future
Resolves classification & regression problems
Training
Inputs → Outputs
Unsupervised Learning
Machine understands the data (Identifies patterns/ structures)
Evaluation is qualitative or indirect
Does not predict / find anything specific
Inputs → Outputs
Reinforcement Learning
An approach to AI
Reward based learning
Learning from +ve & -ve reinforcement
Machine learns how to act in a certain environment
To maximize rewards
Rewards
Inputs → Outputs

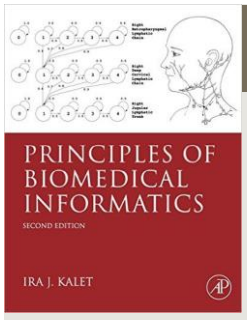

Low granularity
Low cost/patient
Population Health-Registry
Regional incidence
Environmental influences
requires all patients regionally
Quality Reporting
Disparities of care
Practice quality improvement
Safety
requires all patients from participating practices
Decision Support
Outcome/Toxicity prediction
Individualized treatments
Data-driven quality control
Treatment adaptation
large random sample of patients
Research
Biological questions
Clinical trials
statistically sufficient patients
Big data opportunity
Number of patients
High granularity
High cost/patient



Kagadis *et al.*: Medical physicists and health care applications of informatics







Ira Kalet, PhD

Remembering Ira Kalet, 1944-2015

Retired CSE adjunct professor Ira Kalet passed away last night after a long battle with cancer.

Ira joined the University of Washington in 1978 in the then newly formed Department of Radiation Oncology. Subsequently he held adjunct appointments in Computer Science & Engineering, Bioengineering, and Biological Structure, and a joint appointment in Medical Education (now the Department of Biomedical Informatics and Medical Education).

SPECIAL ARTICLE

Technology for Innovation in Radiation Oncology

Indrin J. Chetty, PhD,* Mary K. Martel, PhD,† David A. Jaffray, PhD,‡

1. **Integrating radiation oncology databases across the discipline will facilitate science and elevate the quality of care** (45). The creation of a Virtual Clinical Trials Group that enables federated databases at different institutions for conducting cooperative research is a consideration. Sharing practices and outcomes will permit high mean and tight variance in clinical practice and will improve quality (46).

2. **Tools need to be created and made available for patients and physicians to discuss treatment options,** as recommended by the Patient-Centered Outcome Research Institution. Such an approach will drive the development of metatreatment planning systems, in which one prescribes an outcome, not a treatment (eg specification of a 95% local control rate at 5 years with 5% grade 3 or more dyspnea) (6, 47). This could also be expanded beyond radiation oncology.

3. **Expertise in the informatics domain among radiation oncology professionals needs to be developed** (6). The most suitable candidates with the appropriate skill sets and multidisciplinary knowledge to succeed in this space are likely medical physicists or physicians with strong computational backgrounds. Training grants for developing programs for oncology informatics will provide these individuals with the knowledge needed to support informatics research initiatives.

4. **Informatics tools need to be developed to support the monitoring of the quality of oncology care at the point(s) of delivery** (48). Real world–based evidence approaches are emerging in other domains and will also benefit the field of radiation oncology. The often-quoted statements that 5% differences in dose result in significant changes in tumor control and normal tissue complication probabilities will be reinforced or challenged through collecting and sharing data from the entire clinical process.

NIH **NATIONAL CANCER INSTITUTE**
Informatics Technology for Cancer Research

Funding Opportunities

ITCR has issued four Funding Opportunity Announcements aimed at successive stages of informatics technology development.

Algorithm Development

PAR-15-334 Development of Innovative Informatics Methods and Algorithms for Cancer Research and Management (R21)

Prototyping & Hardening

PAR-15-332 Early-Stage Development of Informatics Technologies for Cancer Research and Management (U01)

Enhancement & Dissemination

PAR-15-331 Advanced Development of Informatics Technologies for Cancer Research and Management (U24)

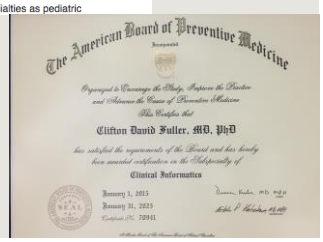
Sustainment

PAR-15-333 Sustained Support for Informatics Resources for Cancer Research and Management (U24)

Clinical Informatics Becomes a Board-certified Medical Subspecialty Following ABMS Vote

Thursday, September 22, 2011
AMIA to offer prep courses for clinicians who sit for Board Exam

Washington, DC—Today, AMIA—the association for informatics professionals—announces the success of a multi-year initiative to elevate clinical informatics to an American Board of Medical Specialties (ABMS) subspecialty certified by an examination administered by the American Board of Preventive Medicine and available to physicians who have primary specialty certification through the American Board of Medical Specialties. Joining such subspecialties as pediatric


The American Board of Preventive Medicine

Clifton David Fuller, MD, PhD

has satisfied the requirements of the Board and has hereby been awarded certification in the Subspecialty of

Clinical Informatics

January 1, 2015
January 31, 2025

Fellowship in Clinical Informatics: Radiation Oncology Track

Fellowship in Clinical Informatics: Radiation Oncology Track

Clinical informatics is the subspecialty of all medical specialties that transforms health care by analyzing, designing, implementing, and evaluating information and communication systems to improve patient care, enhance access to care, advance individual and population health outcomes, and strengthen the clinician-patient relationship.

OHSU School of Medicine
Radiation Medicine

Health Informatics via Machine Learning for the Clinical Management of Patients

D. A. Clifton, K. E. Niehaus, P. Charlton, G. W. Colopy

Yearb Med Inform 2015;10:38-43
http://dx.doi.org/10.15265/IY-2015-014
Published online August 13, 2015

4 Conclusions

We conclude by emphasising that the field of health informatics systems based on machine learning, drawing on disparate datatypes from the ICU, the wider hospital, and from (potentially very complex) EHR data, is in its infancy. While the majority of hospitals in the developed world have implemented EHR systems of some kind, the integrated use of the large quantities of data that arise from such systems is not employed at scale. This

Review of Developments in Electronic, Clinical Data Collection, and Documentation Systems over the Last Decade – Are We Ready fo Big Data in Routine Health Care?

Kerstin A. Kessel[1,2] and Stephanie E. Combs[1,2]*

[1]Department of Radiation Oncology, Technische Universität München, Munich, Germany, [2]Institute of Innovative Radiotherapy (iRT), Helmholtz Zentrum München, Neuherberg, Germany

frontiers in Oncology
REVIEW
published: 30 March 2016
doi: 10.3389/fonc.2016.00075

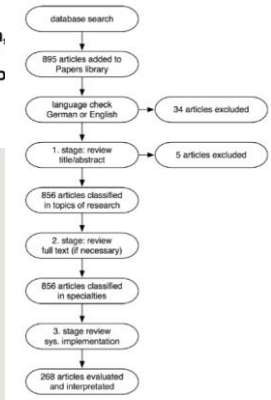


FIGURE 1 | Flow chart of the review methodology.

The curse of clinical implementation: From n=895 to n=5

TABLE 2 | Topics of articles.

Topic	No. of articles
System use	469
• For clinical trial or analysis	370
• For clinical routine	99
System implementation	268
System comparisons with paper-based standard or other systems	24
System review, recommendations, and issues	95
Not assigned	39
	N = 895

TABLE 3 | Articles with further processing strategies and approaches of collected data.

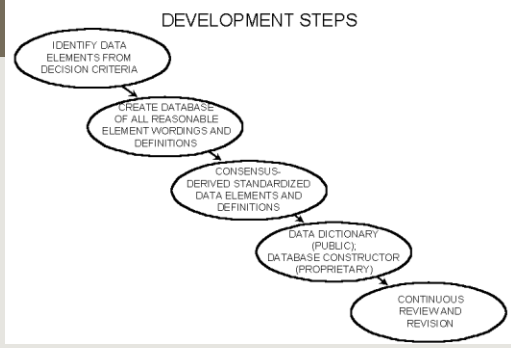
Reference	Year	Summary
Brown et al. (16)	2007	Analysis tools connected to data management system for quantitative image analysis in metastatic lung cancer patients; automatic nodule detection and segmentation for CAD evaluation; communication standards used: DICOM
Carey et al. (17)	2012	Analysis tools used on imaging files stored in database in lung cancer patients; manual image analysis; no communication standardization mentioned
Haak et al. (11, 18)	2014	Analysis tools connected to EDC system for automatic image and biosignal analysis; communication standards used: web services, COM, SOAP, SFTP, HTTP
Kessel et al. (12, 19)	2012	Analysis tools connected to documentation database via SQL interface; semiautomatic CT image registration and segmentation of pancreatic cancer patients, as well as dose calculation of radiation plans; communication standards used: HL7, DICOM, https
Cayurt et al. (20)	2010	Analysis tools used on local copies of neuroimaging data after query and download from the data management system; results are transferred back via web services; communication standards used: web services, SOAP, DICOM, https

CAD, computer-aided diagnosis; DICOM, digital imaging and communications in medicine; EDC, electronic data capture; HL7, health level 7; SQL, structured query language; COM, object data model; SOAP, simple object access protocol.

Kessel and Combs Critique

It is not the lack of technology or tools that keep "the health revolution" from coming, but the lack of expertise, specifications, and concepts (4). One of the most common weaknesses found is the lack of standardization. Most researchers create an individual in-house solution without considering communication standards, such as DICOM, HL7, https, and html (37). These solutions work only in their own environment and are tailored to meet their requirements. This might be necessary up to a certain level, as already stated that there is no "one-size-fits-all" solution for documentation of clinical trials, research data, or patient data

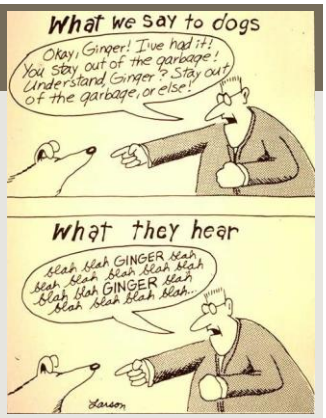
FIGURE 3. DATA STANDARDS/DICTIONARY



- K. Hammermeister, MD (http://www.uni-mainz.de/FB/Medizin/Kardiologie/incis/Data/p4_1.htm)

The "ontology" problem

- Without common terminology, content is obscured…and we may not be aware of it!



Standards and ontologies

- In informatics, a standard is voluntary consensus developed or adopted by voluntary bodies
  - Can also include
    - definitions
    - a classification system or components,
    - a series of procedures to follow,
    - specifications for sizes or dimensions, materials, performance, design or operations.
    - quality measures or describe amounts of materials.
    - process, a system, a service, a practice, or a specific product.
- In computer science and information science, an ontology is a formal naming and definition of the types, properties, and interrelationships of the entities that really or fundamentally exist for a particular domain of discourse.
- Example: DICOM-RT is a standard, TG-263 OAR names are an ontology.
- Functionally speaking, a relational standard=an ontology.

EXAMPLE: Can we just name OARs the same?

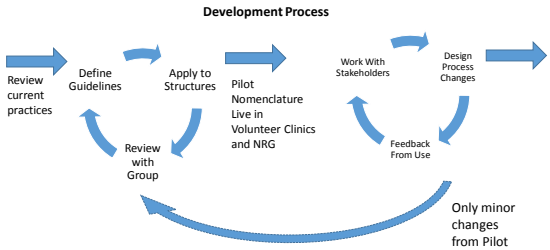
Let's start by trying to fix the standardization problems for DVH data

TG 263 - Standardizing Nomenclature for Radiation Therapy

- group of 57 stake holders
- domestic and international groups
- representing a broad range of perspectives

Roles	Professional Societies	Clinic Types	Specialty Groups
Physician	ASTRO	Academic	IHE-RO
Physicist	AAPM	Community	Dicom Working Group
Vendor	ESTRO	Large Practice	NRG
Dosimetry		Small Practice	IROC

Slide courtesy of Chuck Mayo (U. Mich.)



Slide courtesy of Chuck Mayo (U. Mich.)

Task Group findings are in parent committee review process

- Guidelines
  - Target Structures
    - Standardized rule based approach (10)
    - Addresses primary issues and expandable
  - Non-Target Structures
    - Rule based approach (15) with a few concessions
    - Specific listing of 736 defined structures
- DVH Nomenclature

Slide courtesy of Chuck Mayo (U. Mich.)

TABLE 5.3 Base Names for Organs at Risk

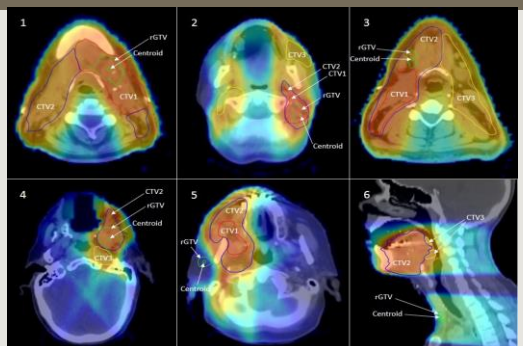
Structure Name	Paired?	Structure Name	Paired?
ANAL_CANAL		MAIN_BRONC	
BLADDER		OPTIC_NRV	_L/_R
BRAC_PLX	_L/_R	ORAL_CAVITY	
BRAIN		OVARY	_L/_R
BRAINSTEM		PAROTID	_L/_R
BREAST	_L/_R	PENILE_BULB	
BRONC_TREE		PERINEUM	
CARINA		PHARYNX	
CAUDA_ EQUINA		PITUITARY	
CEREBELLUM	_L/_R	PROSTATE	
CEREBRUM	_L/_R	RECTUM	
CHIASM		RETINA	_L/_R
CN_VII	_L/_R	RIB	
CN_VIII	_L/_R	SACRUM	
COCHLEA	_L/_R	SEM_VES	
CORNEA	_L/_R	SKIN	
DUODENUM		SM_BOWEL	
EAR_MID	_L/_R	SPINAL_CORD	
EAR_EXT	_L/_R	STOMACH	
ESOPHAGUS		SUBMND_SALV	_L/_R
FEMUR	_L/_R	TEMP_LOBE	_L/_R
GLOBE	_L/_R	TESTIS	_L/_R
GLOTTIS		THYROID	
GREAT_VESS		TM_JOINT	_L/_R
HEART		TONGUE	
KIDNEY	_L/_R	TRACHEA	
LG_BOWEL		URETHRA	
LARYNX		VULVA	
LAC_GL	_L/_R		
LENS	_L/_R		
LIPS			
LIVER			
LUNG			
MANDIBLE			

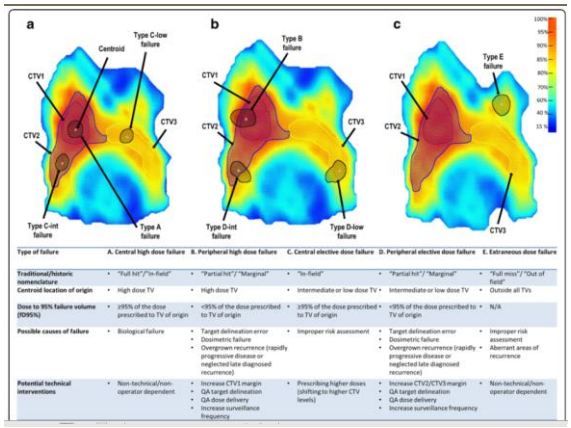
Source: From Bosch, W. R., *Uniform Tissue Names for Use in RTOG Advanced Technology Clinical Trials*, August 19, 2009. Available at http://atc.wustl.edu. Accessed 1 December 2009.

Note: Where paired organs are indicated, laterality is indicated by appending "_L" or "_R" to the base name. Geometric extensions of these structures are indicated by appending "_PRVxx" where *xx* is the nominal margin (mm) used to extend the structure.

We're just now agreeing on the ontology of structure names!!!

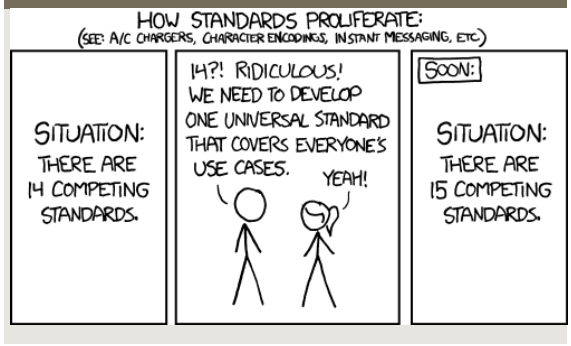
Ontology issue example: "How do you designate recurrence in IMRT era?"





Type of failure	A. Central high dose failure	B. Peripheral high dose failure	C. Central elective dose failure	D. Peripheral elective dose failure	E. Extraneous dose failure
Traditional/historic nomenclature	• "Full hit"/"In-field"	• "Partial hit"/ "Marginal"	• "In-field"	• "Partial hit"/ "Marginal"	• "Full miss"/ "Out of field"
Centroid location of origin	• High dose TV	• High dose TV	• Intermediate or low dose TV	• Intermediate or low dose TV	• Outside all TVs
Dose to 95% failure volume (D95%)	• ≥95% of the dose prescribed to TV of origin	• <95% of the dose prescribed to TV of origin	• ≥95% of the dose prescribed to TV of origin	• <95% of the dose prescribed to TV of origin	• N/A
Possible causes of failure	• Biological failure	• Target delineation error • Dosimetric failure • Overgrown recurrence (rapidly progressive disease or neglected late diagnosed recurrence)	• Improper risk assessment	• Target delineation error • Dosimetric failure • Overgrown recurrence (rapidly progressive disease or neglected late diagnosed recurrence)	• Improper risk assessment • Aberrant areas of recurrence
Potential technical interventions	• Non-technical/non-operator dependent	• Increase CTV1 margin • QA target delineation • QA dose delivery • Increase surveillance frequency	• Prescribing higher doses (shifting to higher CTV levels)	• Increase CTV2/CTV3 margin • QA target delineation • QA dose delivery • Increase surveillance frequency	• Non-technical/non-operator dependent

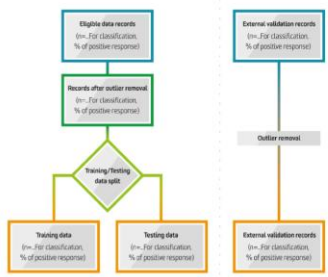
The Standards Problem/Opportunity



OURNAL OF MEDICAL INTERNET RESEARCH
oi:10.2196/jmir.5870

Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View

Figure 2. Information flow in the predictive modelling process.



Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View

doi:10.2196/jmir.5870

7	Prepare data for model building	Identify relevant data sources and quote the ethics approval number for data access. State the inclusion and exclusion criteria for data. Describe the time span of data and the sample or cohort size. Define the observational units on which the response variable and predictor variables are defined. Define the predictor variables. Extra caution is needed to prevent information leakage from the response variable to predictor variables.[c] Describe the data preprocessing performed, including data cleaning and transformation. Remove outliers with impossible or extreme responses; state any criteria used for outlier removal. State how missing values were handled. Describe the basic statistics of the dataset, particularly of the response variable. These include the ratio of positive to negative classes for a classification problem and the distribution of the response variable for regression problem. Define the model validation strategies. Internal validation is the minimum requirement; external validation should also be performed whenever possible. Specify the internal validation strategy. Common methods include random split, time-based split, and patient-based split. Define the validation metrics. For regression problems, the normalized root-mean-square error should be used. For classification problems, the metrics should include sensitivity, specificity, positive predictive value, negative predictive value, area under the ROC[d] curve, and calibration plot [19].[e] For retrospective studies, split the data into a derivation set and a validation set. For prospective studies, define the starting time for validation data collection.

Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View

doi:10.2196/jmir.5870

8	Build the predictive model	Identify independent variables that predominantly take a single value (eg, being zero 99% of the time). Identify and remove redundant independent variables. Identify the independent variables that may suffer from the perfect separation problem.[f] Report the number of independent variables, the number of positive examples, and the number of negative examples. Assess whether sufficient data are available for a good fit of the model. In particular, for classification, there should be a sufficient number of observations in both positive and negative classes. Determine a set of candidate modeling techniques (eg, logistic regression, random forest, or deep learning). If only one type of model was used, justify the decision for using that model.[g] Define the performance metrics to select the best model. Specify the model selection strategy. Common methods include K-fold validation or bootstrap to estimate the lost function on a grid of candidate parameter values. For K-fold validation, proper stratification by the response variable is needed.[h] For model selection, include discussion on (1) balance between model accuracy and model simplicity or interpretability, and (2) the familiarity with the modeling techniques of the end user.[i]

Guidelines for Developing and Reporting Machine Learning Predictive Models in Biomedical Research: A Multidisciplinary View

doi:10.2196/jmir.5870

Table 5. Items to include when reporting predictive models in biomedical research: discussion section.

Item number	Topic	Checklist item
10	Clinical implications	Report the clinical implications derived from the obtained predictive performance. For example, report the dollar amount that could be saved with better prediction. How many patients could benefit from a care model leveraging the model prediction? And to what extent?
11	Limitations of the model	Discuss the following potential limitations: • Assumed input and output data format • Potential pitfalls in interpreting the model[a] • Potential bias of the data used in modeling • Generalizability of the data
12	Unexpected results during the experiments	Report unexpected signs of coefficients, indicating collinearity or complex interaction between predictor variables[a]

Baby steps…

- Why is clinical implementation so hard?
  - Nature of Big Data
    - Sparseness of Big Data
    - Quality assurance
  - Nature of machine learning
    - Algorithms only ”know what you tell them, or ask what you ask them to find”
    - Interpretability of complex datasets harder for humans than for algorithms sometimes

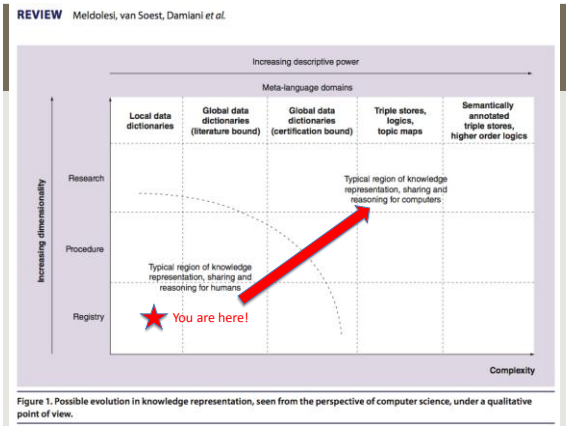


Figure 1. Possible evolution in knowledge representation, seen from the perspective of computer science, under a qualitative point of view.

32 Flavors of machine learning algorithms…

Table 2
Benefits and limitations of different machine learning algorithms.

Algorithm	Advantages	Limitations
Decision Tree	• Easy to understand • Fast	• Classes must be mutually exclusive • Results depend on the order of attribute selection • Risk of overly complex decision trees
Naive Bayesian	• Easy to understand • Fast • No effect of order on training	• Variables must be statistically independent • Numeric attributes must follow a normal distribution • Classes must be mutually exclusive • Less accurate
k-nearest Neighbors	• Fast and simple • Tolerant of noise and missing values in data • Can be used for non-linear classification • Can be used for both regression and classification	• Variables with similar attributes will be sorted in the same class • All attributes are equally relevant • Requires considerable computer power as number of variables increases
Support Vector Machine	• Robust model • Limits the risk of error	• Slow training • Risk of overfitting
Artificial Neural Network and Deep Learning	• Can be used to model non-linear relations • Tolerant of noise and missing values in data • Can be used for classification or regression • Can be easily updated with new data	• Output model is difficult to understand • Output model is difficult to understand (« black-box ») • Risk of overfitting • Requires a lot of computer power • Requires experimentation to find the optimal network structure

Remember, a HUMAN PHYSICIAN has to understand the ML output…

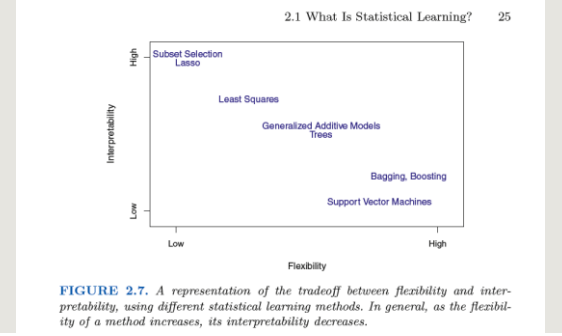
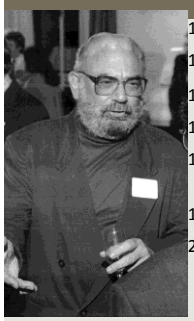


FIGURE 2.7. *A representation of the tradeoff between flexibility and interpretability, using different statistical learning methods. In general, as the flexibility of a method increases, its interpretability decreases.*

Leo Breiman, 1928 - 2005



1954 PhD Berkeley (mathematics)

1960 -1967 UCLA (mathematics)

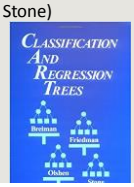
1969 -1982 Consultant

1982 - 1993 Berkeley (statistics)

1984 "Classification & Regression Trees" (with Friedman, Olshen, Stone)

1996 "Bagging"

2001 "Random Forests"



RECURSIVE PARTITIONING ANALYSIS (RPA) OF PROGNOSTIC FACTORS IN THREE RADIATION THERAPY ONCOLOGY GROUP (RTOG) BRAIN METASTASES TRIALS

LAURIE GASPAR, M.D.,* CHARLES SCOTT, M.S.,† MARVIN ROTMAN, M.D.,‡ SUCHA ASBELL, M.D.,§ THEODORE PHILLIPS, M.D.,¶ TODD WASSERMAN, M.D.,# W. GILLIES MCKENNA, M.D., Ph.D.** AND ROGER BYHARDT, M.D.††

746 I. J. Radiation Oncology ● Biology ● Physics Volume 37, Number 4, 1997

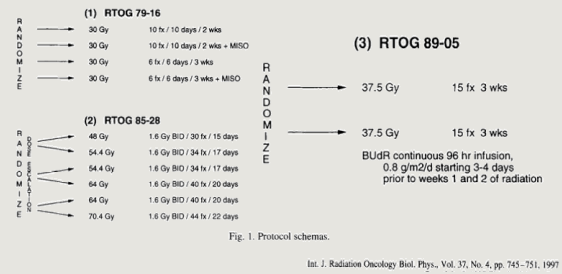


Fig. 1. Protocol schemas.

Int. J. Radiation Oncology Biol. Phys., Vol. 37, No. 4, pp. 745–751, 1997

Radiation oncologists primary use of "machine learning" has been "recursive partitioning analysis" (RPA)

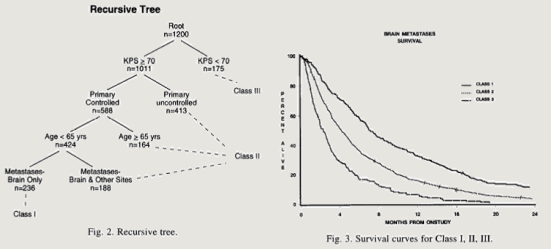


Fig. 2. Recursive tree.

Fig. 3. Survival curves for Class I, II, III.

Int. J. Radiation Oncology Biol. Phys., Vol. 37, No. 4, pp. 745–751, 1997

Result: Rad Onc loves RPA!!

- RPA is a "brand name"
  - Classification and regression trees
  - Decision trees
  - Tree-based analysis
- We got into the game early…
  - But we have a long ways to go in machine learning

Example: ML for toxicity/dose response

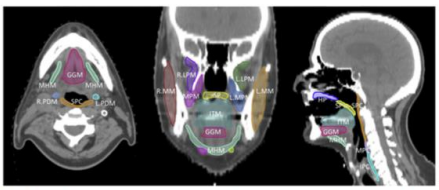
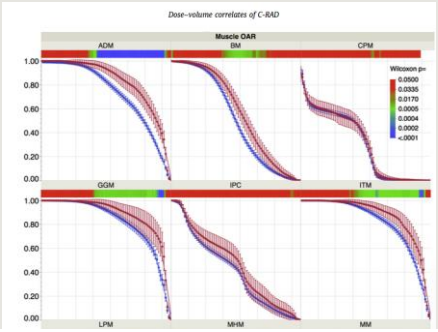


Fig. 1. Exemplar swallow-related ROI. Axial, coronal, and sagittal images of the contoured segments. *Abbreviations:* GGM – genioglossus muscle; HP – hard palate; IPC – inferior pharyngeal constrictor; ITM – intrinsic tongue muscles; LPM – lateral pterygoid muscle; MHM – mylo/geniohyoid complex; MM – masseter muscle; MPM – medial pterygoid muscle; PDM – posterior digastric muscle; SP – soft palate; SPC – superior pharyngeal constrictor; R.-right, L.-left.

Radiotherapy and Oncology 118 (2016) 304–314

RPA for thresholding doses that indicate toxicity…



RPA-based method

Table 2
Univariate RPA-derived ROI specific dose–volume thresholds. LogWorth represents the negative

Recursive partitioning analysis							
Muscle OAR	V-level	Percent-threshold (%)	ROC AUC cohort (test)	ROC AUC holdback (verification)	LogWorth	*p*-Value	SS
ADM	60	79	0.68	0.60	5.95	<.0001	**
BM	35	65.8	0.65	0.57	1.09	0.0815	n.s.
CPM	45	0.35	0.64	0.51	1.00	0.0998	n.s.
GGM	35	98.9	0.70	0.55	2.74	0.0018	**
IPC	70	98.2	0.60	0.51	1.08	0.0831	n.s.
ITM	47	99.9	0.67	0.44	2.83	0.0015	*
LPM	66	13.1	0.53	0.35	1.07	0.0860	n.s.
LRX	63	1	0.61	0.47	0.89	0.1274	n.s.
MHM	69	17.5	0.74	0.64	6.77	<.0001	**
MM	66	4.4	0.61	0.53	0.88	0.1314	n.s.
MPC	49	99.9	0.63	0.54	0.17	0.6825	n.s.
MPM	70	1	0.59	0.45	3.31	0.0005	*
PDM	69	13.5	0.60	0.48	0.15	0.7070	n.s.
PGM	65	68.9	0.62	0.49	0.24	0.5732	n.s.
SPC	70	6.35	0.68	0.47	5.09	<.0001	**

* Statistically significant at $P < 0.05$.
** Statistically significant after Bonferroni correction.

RPA-based method

Model	Model effects	Model goodness of fit*	Whole model LogWorth †	Whole model p-value	ROC AUC Cohort (test)	ROC AUC Holdback (verification)	Model BIC [ΔBIC]	Evidence Grade §	Effect likelihood LogWorth †	Effect likelihood p-value	Odds Ratio (95% CI)
Continuous	MHM V69	0.9999	6.94	<.0001	0.835	0.812	124.25 [0]	BIC_{minimum} (reference)	5.19	<.0001	1.09 (1.03–1.16) ‡
	Age								2.56	0.0027	1.03 (1.02–1.05) ‡
Binary	MHM V69 >79.5%	0.9992	8.76	<.0001	0.825	0.864	130.53 [6.28] §	Strong	7.62	<.0001	6.8×10^{7} (12.63–∞) ◊
	Age >62								1.37	0.043	2.57 (1.03–6.53) ◊

MDACC RPA-based method showed Myelohyoid muscle V69 and age were best predictors…

Development of a multivariable normal tissue complication probability (NTCP) model for tube feeding dependence after curative radiotherapy/chemo-radiotherapy in head and neck cancer

Kim Wopken [a,*], Hendrik P. Bijl [a], Arjen van der Schaaf [a], Hans Paul van der Laan [a], Olga Chouvalova [a], Roel J.H.M. Steenbakkers [a], Patricia Doornaert [b], Ben J. Slotman [b], Sjoukje F. Oosting [c], Miranda E.M.C. Christianen [a], Bernard F.A.M. van der Laan [d], Jan L.N. Roodenburg [e], C. René Leemans [f], Irma M. Verdonck-de Leeuw [f], Johannes A. Langendijk [a]

Table 3
Results of the LASSO analysis with tube feeding dependence at 6 months ($TUBE_{M6}$) as primary endpoint.

Variable	OR	OR 95% CI	p-Value
T-classification			
Tis-T2	1.00		
T3-T4	1.53	(1.17–2.06)	<0.001
Baseline weight loss			
No weight loss	1.00		
Moderate weight loss (1–10%)	2.58	(2.01–3.19)	<0.001
Severe weight loss (>10%)	5.08	(3.32–7.30)	<0.001
Treatment modality			
Conventional fractionation	1.00		
Radiotherapy + cetuximab	1.74	(1.50–2.01)	<0.001
Accelerated fractionation	3.33	(2.40–4.53)	<0.001
Chemoradiation	6.73	(4.00–10.98)	<0.001
Dosimetric variables			
PCM superior mean dose (Gy)	1.07	(1.04–1.09)	<0.001
PCM inferior mean dose (Gy)	1.03	(1.01–1.05)	0.006
Contralateral parotid mean dose (Gy)	1.01	(1.00–1.02)	0.14
Cricopharyngeal muscle mean dose (Gy)	1.02	(1.01–1.03)	0.004

Abbreviations: OR, odds ratio; CI, confidence interval; PCM, pharyngeal constrictor muscle.

UMCG Lasso method showed pharyngeal constrictor muscle and cricopharyngeus were best predictors…

Radiotherapy and Oncology 113 (2014) 95–101

Who was right?

- Answer: Everyone!
  - Solid methods/different datasets
  - Dose variation was due to tumor location/site, and dose to OARs, not “magic dose threshold”
  - MDACC all oropharynx, block over cricopharyngeus m., posterior constrictors m. with low variability as they were partially covered by RP node CTVs.
  - UMCG mostly larynx, little dose to myelohyoid m.

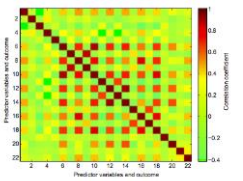


Fig. 1. Image plot of the correlation matrix of the xerostomia data. Pearson's correlation coefficients among predictor variables are shown in different colors from green to red. Red indicates high correlation between two variables. Here variables are named by their index, and their description is listed in Table.

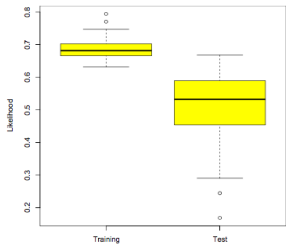
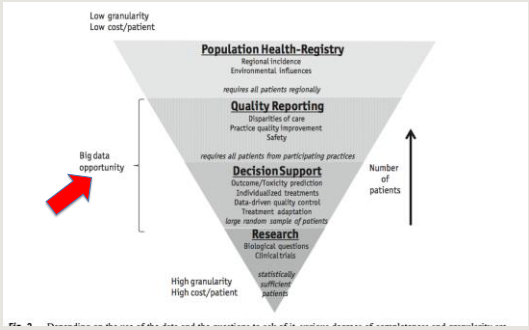


Fig. 2. A box plot of the fit performance in training sets an prediction performance in test sets for xerostomia data of all variable model is shown. The median of the likelihood is indi cated by the center line, and the first and third quartiles are th edges of the box area. The extreme values (within 1.5 times th interquartile range from the upper or lower quartile) are the end of the lines extending from the interquartile range. Points a a distance from the median greater than 1.5 times the interquartil range are plotted individually (circles).

Xu et al. Int J Radiation Oncol Biol Phys, Vol. 82, No. 4, pp. e677–e684, 2012

Problem/opportunity: We don't have accessible aggregate data…



Peeling the onion…

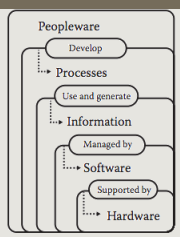


FIGURE 7.1 Modern radiation oncology practice comprises five major components: peopleware, processes, information, software, and hardware. Successful development, implementation, and use of IS and IT applications requires proper consideration of the hierarchy as well as synergy among each of the components.

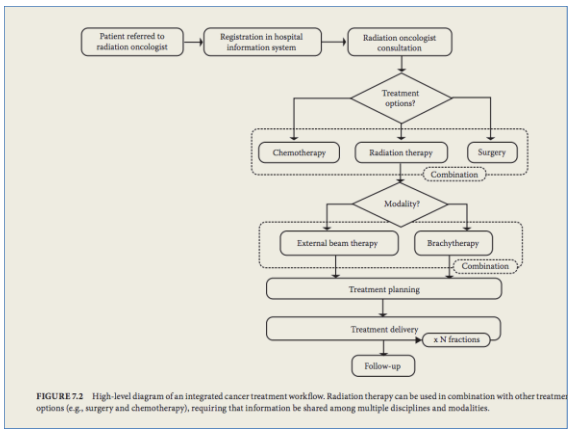


FIGURE 7.2 High-level diagram of an integrated cancer treatment workflow. Radiation therapy can be used in combination with other treatmer options (e.g., surgery and chemotherapy), requiring that information be shared among multiple disciplines and modalities.

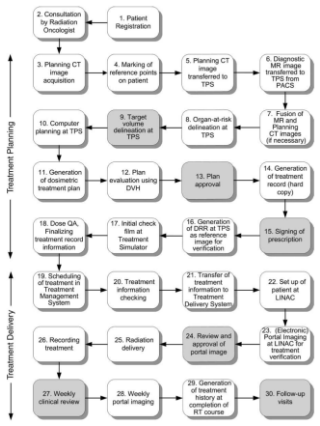
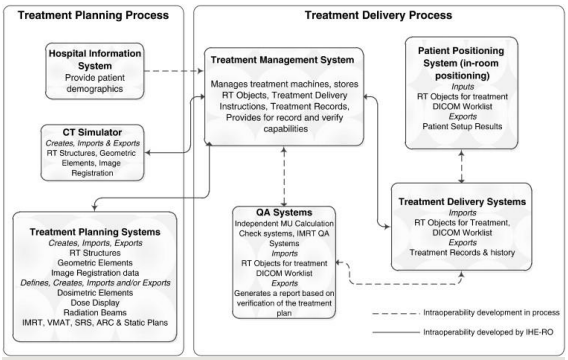


Figure 1. Chart illustrates the radiation therapy work flow for a patient with prostate cancer. The work flow consists of two stages, treatment planning (steps 1–18) and treatment delivery (steps 19–30), and is broken down into steps that can be used to evaluate the efficiency of the system once it is implemented. The work flow starts when the radiation oncologist decides at consultation that the patient is to receive radiation treatment (step 2). Gray boxes indicate steps that are performed by the radiation oncologist. Of these steps, steps 9 and 13 are incorporated into the treatment planning system (TPS) in most cases, but steps 15, 24, 27, and 30 either require hard-copy records or radiation therapy information that may not be readily accessible. *DRR* = digitally reconstructed radiograph, *DVH* = dose-volume histogram, *LINAC* = linear accelerator, *QA* = quality assurance.



GeneMed: An Informatics Hub for the Coordination of Next-Generation Sequencing Studies that Support Precision Oncology Clinical Trials

Yingdong Zhao[1,*], Eric C. Polley[1,*], Ming-Chung Li[2], Chih-Jian Lih[2], Alida Palmisano[1], David J. Sims[3], Lawrence V. Rubinstein[1], Barbara A. Conley[3], Alice P. Chen[4], P. Mickey Williams[2], Shivaani Kummar[4], James H. Doroshow[4] and Richard M. Simon[1]

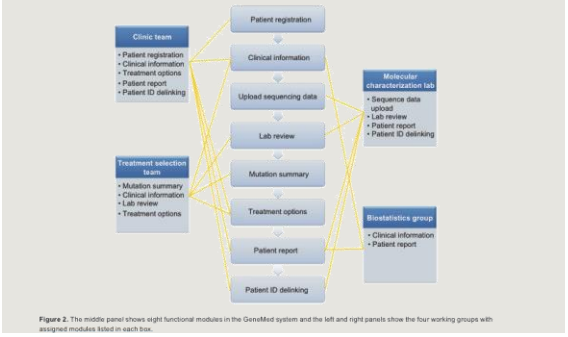
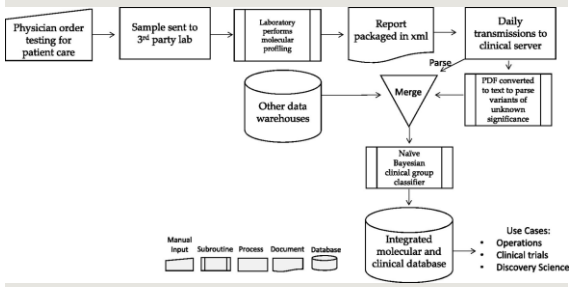


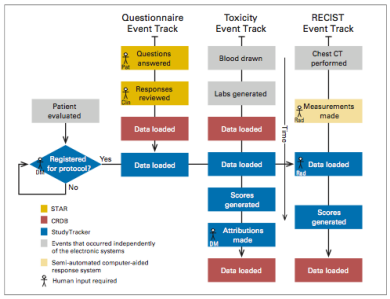
Figure 2. The middle panel shows eight functional modules in the GeneMed system and the left and right panels show the four working groups with assigned modules listed in each box.

Biospecimen/pathology informatics critical



Harnessing Technology to Improve Clinical Trials: Study of Real-Time Informatics to Collect Data, Toxicities, Image Response Assessments, and Patient-Reported Outcomes in a Phase II Clinical Trial

M. Catherine Pietanza, Ethan M. Basch, Alex Lash, Lawrence H. Schwartz, Michelle S. Ginsberg, Binsheng Zhao, Marwan Shaury, Mary Shaw, Lauren J. Rogak, Manda Wilson, Aaron Gabow, Marcia Latif, Kai-Hsiang Lin, Qinfei Wu, Samantha L. Kass, Claire P. Miller, Leslie Tyson, Dyana K. Sumner, Alison Berkowitz-Hergianto, Camelia S. Sima, and Mark G. Kris



An Open-Source and Open-Access Cancer Biomedical Informatics Grid Annotation and Image Markup Template Builder[1]

TEACHING POINTS See last page

Puttanaah Mongkolwat, PhD • David S. Channin, MD • Vladimir Kleper
Daniel L. Rubin, MD

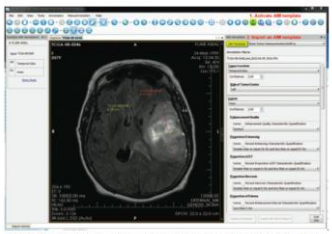


Figure 2. Screen shot from the open-source ClearCanvas workstation shows how AIM templates can be used to open AIM template documents, select the desired templates, and fill out template-based forms to create AIM objects. The template-based forms ask questions with a constrained range of answers and may require the user to perform measurements or create markups. The AIM template functionality can be activated by clicking on the highlighted AIM button on the workstation toolbar (shown in yellow, with red text). AIM templates can be imported using the highlighted "AIM Template" button (shown in yellow, with red text).

The Cancer Journal • Volume 17, Number 4, July/August 2011

TABLE 1. Producers of Information in Clinical Trials

Producers of Information in Clinical Trials Are as Follows:
A. Pre-enrollment
• The authors of the protocol
• The IRBs that approve it for local participation
• The lead institution, sponsor, or CRO that creates the CRFs for EDC
• The individuals (coordinators/clinicians, individual investigators, and budgeting/billing administrators) that interpret data for that protocol into local systems, including scheduling, budgeting, reviews, and departmental and institutional reporting.
• Clinical trial budgeting and contract management staff
B. Intraenrollment
• The subject/patient
• The clinician, both in CRFs and in source documentation (the medical record—on paper or electronic)
• The laboratory (hematology, chemistry, genomics, microbiology, pathology) and radiology departments and systems
• Research monitors (queries and reviews)
• Internal and external regulatory reviewers
• Internal and external auditors

Data QA is science QA

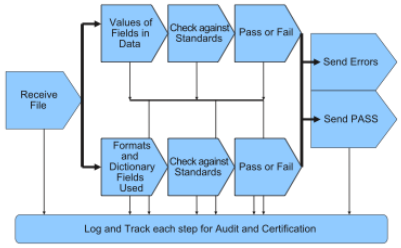


FIGURE 5. Validation for format, fields, and values against standards: a simple configuration for standards designers.

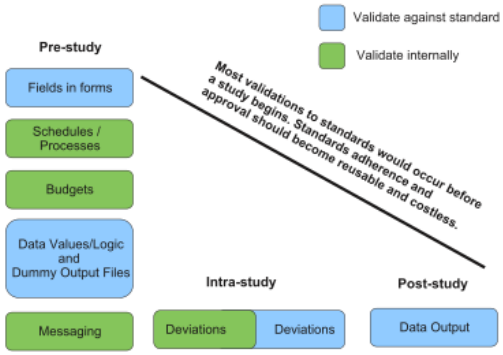
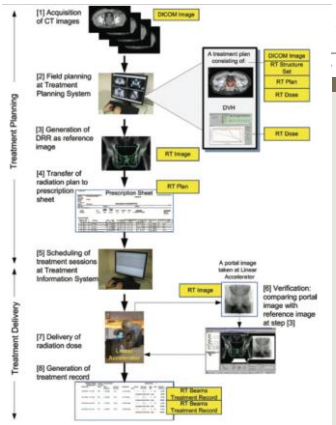


FIGURE 6. Possible automated validations: need and timing.



DICOM-RT and Its Utilization in Radiation Therapy[1]

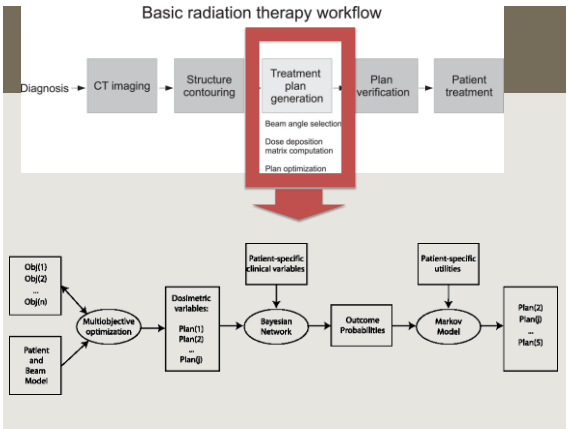
Maria Y. Y. Law, PhD • Brent Liu, PhD

Figure 3. Chart illustrates radiation therapy work flow. Yellow boxes indicate the DICOM-RT objects that could be generated within the work flow. A radiation therapy treatment plan (step 2) with radiation dose distribution involves the superposition of the radiation therapy objects RT Plan, RT Structure Set, and RT Dose on the corresponding set of DICOM computed tomographic (CT) scans according to the coordinates in the DICOM-RT standard. Because the work flow is for external beam therapy, the RT Brachy Treatment Record information object is not shown. *DRR* = digitally reconstructed radiograph, *DVH* = dose-volume histogram.

Table 2
Image Object versus an RT Object IOD Module

Information Entity of DICOM Image Object	Module* (MR Imaging)	Information Entity of DICOM-RT Object	Module* (RT Structure Set)
Patient	Patient	Patient	Patient
Study	General Study	Study	General Study
Series	General Series	Series	RT Series
Frame of Reference	Frame of Reference	...	...
Equipment	General Equipment	Equipment	General Equipment
Image	General Image, Imaging Plane, Image Pixel, MR Image, SOP Common (13)	Structure Set	Structure Set, ROI Contour, RT ROI Observations, SOP Common (13)

*Only mandatory modules are listed. A module groups related information together. For example, the Patient module contains attributes related to the patient, such as Patient's Name, Patient ID, Patient's Birth Date, Patient's Sex, and so on. The modules for the image Information Entity depend on the modality concerned. In this table, the modality of interest is MR imaging; thus, in addition to the General Image module, Information Entity has a specific MR Image module and other related modules. Structure Set is one of the DICOM-RT objects and includes Structure Set, ROI Contour, and RT ROI Observations as its specific modules. IOD = information object definition, ROI = region of interest, SOP = service-object pair.



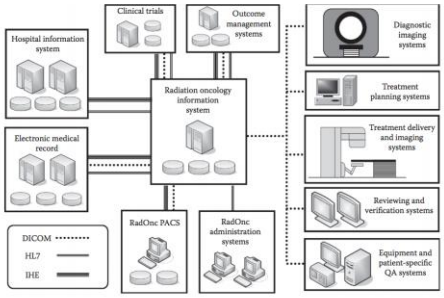
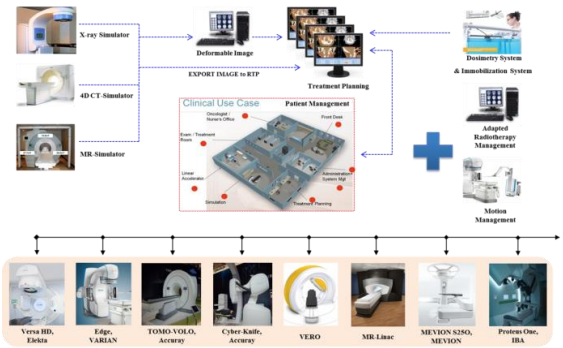


FIGURE 7.5 Hardware components and connectivity infrastructure of the modern radiation oncology practice. Highly available data storage and efficient information flow are highly dependent on having and maintaining reliable, safe, and stable infrastructure and connectivity across all the systems inside and outside the radiation oncology department.



Complex functional imaging

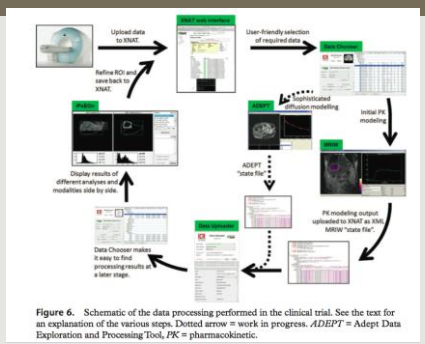


Figure 6. Schematic of the data processing performed in the clinical trial. See the text for an explanation of the various steps. Dotted arrow = work in progress. *ADEPT* = Adept Data Exploration and Processing Tool, *PK* = pharmacokinetic.

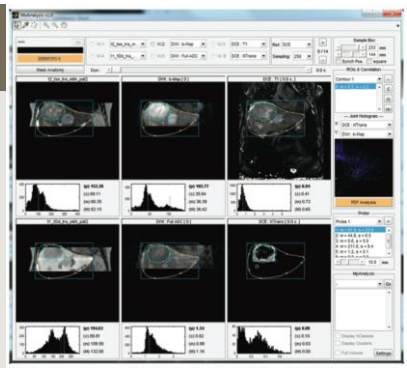


Figure 8. Pharmacokinetic modeling with the tools of the research PACS. Screen shot from iPAEOn shows the following data types: T2-weighted anatomic image (top left), diffusion-weighted image (top center), T1 map from DCE imaging data (calculated in MRIW) (top right), T1-weighted anatomic image (bottom left), ADC map (calculated in iPAEOn) (bottom center), and transfer constant (K^{trans}) map (calculated in MRIW) (bottom right). The iPAEOn application allows imaging data from a variety of sources to be brought into registration and displayed side by side. *DWI* = diffusion-weighted imaging.

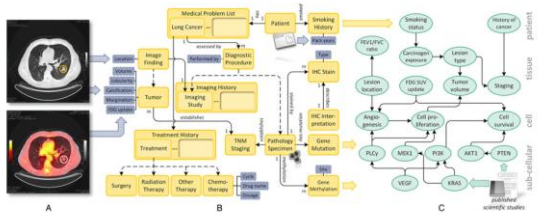
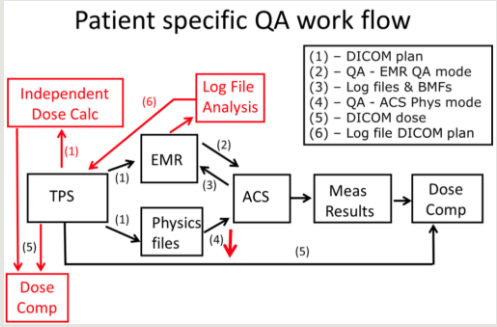
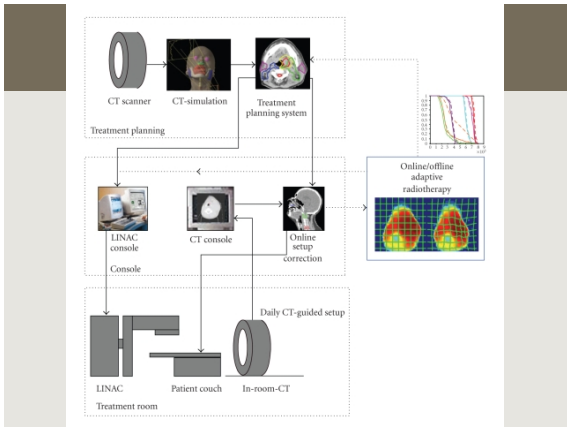
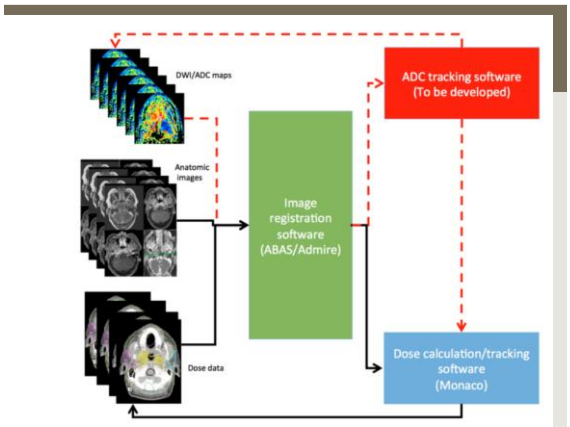


Figure 3 Integrating imaging and clinical information to construct a disease model. (A) Features are computed from normalized imaging studies. In this case, a lung cancer is identified on a CT imaging study (top image), and a corresponding positron emission tomography study (bottom image) is performed to assess the level of tumor activity. (B) An integrated data model creates the basis for combining image data with additional information from the patient's medical record, thereby providing context to the imaging study interpretation. For instance, information on the patient's smoking history, results of thoracic tumor biopsies, and current (and past) treatments lend insight into an individual's status. (C) A multi-scale disease model can be populated from this imaging-based observational database to establish a population-based perspective on the condition (a partial model is shown). Conversely, if a disease model has been vetted for a given population, then the model can be used as a prognostic tool for a given individual, offering tailored recommendations. Knowledge from published scientific studies and other resources can also be embedded within the disease model.

What about non-clinical data that isn't in the PACS, TPS, or EMR, but can affect outcome?







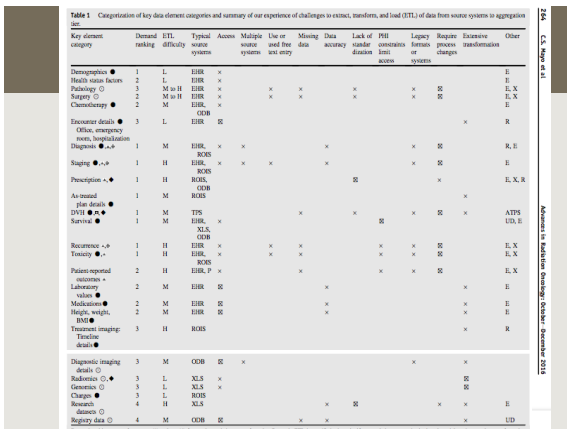


Table 1 Categorization of key data element categories and summary of our experience of challenges to extract, transform, and load (ETL) of data from source systems to aggregation tier.

Key element category	Demand ranking	ETL difficulty	Typical source systems	Access	Multiple source systems	Use or need free text entry	Missing data	Data accuracy	Lack of standardization	PHI constraints limit access	Legacy formats or systems	Require process changes	Extensive transformation	Other
Demographics ●	1	L	EHR	×										E
Health status factors	2	L	EHR	×										E
Pathology ○	3	M to H	EHR	×		×	×		×		×	⊠		E, X
Surgery ○	2	M to H	EHR	×		×	×		×		×	⊠		E, X
Chemotherapy ●	2	M	EHR, ODB	×										E
Encounter details ● Office, emergency room, hospitalization	3	L	EHR	⊠									×	R
Diagnosis ●,▲,◆	1	M	EHR, ROIS	×	×			×			×	⊠		R, E
Staging ●,▲,◆	1	H	EHR, ROIS	×	×	×		×			×	⊠		E
Prescription ▲,◆	1	H	ROIS, ODB						⊠			×		E, X, R
As-treated plan details ●	1	M	ROIS										×	
DVH ●,▲,◆	1	M	TPS				×		×		×	⊠	×	ATPS
Survival ●	1	M	EHR, XLS, ODB	×						⊠				UD, E
Recurrence ▲,◆	1	H	EHR	×		×	×			×	×	⊠		E, X
Toxicity ●,▲	1	H	EHR, ROIS	×		×	×			×	×	⊠		E, X
Patient-reported outcomes ▲	2	H	EHR, P	×			×			×	×	⊠		E, X
Laboratory values ●	2	M	EHR	⊠				×					×	E
Medications ●	2	M	EHR	⊠				×					×	E
Height, weight, BMI ●	2	M	EHR	⊠				×					×	E
Treatment imaging: Timeline details ●	3	H	ROIS										×	R
Diagnostic imaging details ○	3	M	ODB	⊠	×						×		×	
Radiomics ○,◆	3	L	XLS	×									⊠	
Genomics ○	3	L	XLS	×									⊠	
Charges ●	3	L	ROIS											
Research datasets ○	4	H	XLS					×	⊠			×	×	E
Registry data ○	4	M	ODB	⊠			×	×					×	UD

What about outcome/clinical/PRO integration?

Can electronic web-based technology improve quality of life data collection? Analysis of Radiation Therapy Oncology Group 0828

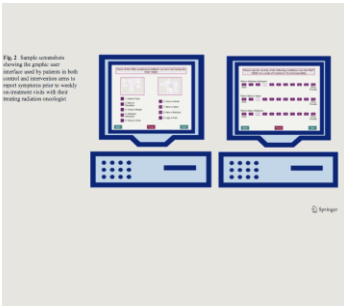
Benjamin Movsas MD [a,*], Daniel Hunt PhD [b], Deborah Watkins-Bruner PhD, RN [c], W. Robert Lee MD, Med [d], Heather Tharpe RN OCN [e], Desiree Goldstein RN MSN [f], Joan Moore RN OCN [g], Ian S. Dayes MD [h], Sara Parise RN OCN [i], Howard Sandler MD [j]

The EPIC QOL compliance rates at baseline, 6 months (the primary endpoint), and 12 months are shown in Table 2. At baseline, the QOL completion rate was 98%. Compared with the 52% 6-month EPIC completion rate using paper forms (in RTOG-0415), the EPIC QOL web-based completion rate at 6 months was 90% (2-sided *P* value < .001). At 1 year, the EPIC QOL web-based compliance rate was 82%. Reasons for noncompliance at 6 months with EPIC QOL via the web-based strategy were patient refusal (2%), patient could not be contacted (2%), or other reason (6%).

Computerized patient-reported symptom assessment in radiotherapy: a pilot randomized, controlled trial

Erik K. Fromme [1] · Emma B. Holliday [2] · Lillian M. Nail [3] · Karen S. Lyons [3] · Michelle R. Hribar [4] · Charles R. Thomas Jr [5]

Support Care Cancer (2016) 24:1897–1906
DOI 10.1007/s00520-015-2983-3



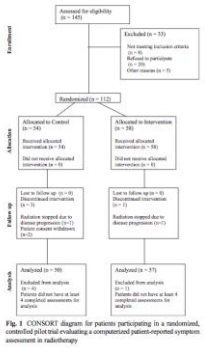


Fig. 1 CONSORT diagram for patients participating in a randomized, controlled pilot trial evaluating a computerized patient-reported symptom assessment in radiotherapy

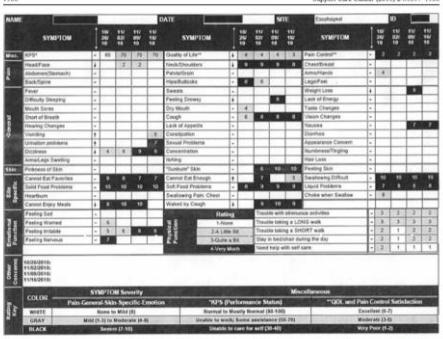


Fig. 3 Sample report showing what was printed and given to clinicians in the intervention groups. Patients received a version with larger font occupying two pages front and back. The *top* shows the survey dates and *columns from the left to the right* show how scores change over time. The *first three rows* show scores for Karnofsky Performance Status (KPS), Quality of Life (QOL), and satisfaction with pain control. The *next three rows* show scores for pain in different anatomic locations. The *next nine rows* show Memorial Symptom Assessment Scale (MSAS) scores. The *following ten rows* show skin toxicity symptoms, other site-specific physical symptoms, and emotional and function scales. The *bottom rows* give a key to interpret the patient-reported outcomes (PROs) for symptoms, KPS, and QOL/pain control satisfaction

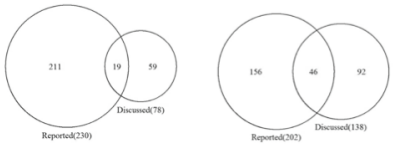


Fig. 4 Venn diagrams and sensitivity accounting to show concordance between patient-reported symptoms and symptoms discussed in weekly office visits for patients in the control and intervention groups. Because this analysis uses data from audio-taped patient-physician interactions, the samples size are the control $n = 26$ and intervention $n = 23$

All Symptoms

	Intervention	Control	P-value
Sensitivity accounting for within subject correlation	23.17%	10.28%	0.03
Sensitivity without accounting for within subject correlation	22.70%	8.30%	

Impact of Statistical Learning Methods on the Predictive Power of Multivariate Normal Tissue Complication Probability Models

Cheng-Jian Xu, Ph.D., Arjen van der Schaaf, Ph.D., Cornelis Schilstra, Ph.D., Johannes A. Langendijk, M.D., Ph.D., and Aart A. van't Veld, Ph.D.

Table Variables in the xerostomia data set

Variable index	Description	Range or limits	Median or frequency	Correlation
1	Chemotherapy*	0, 1	142, 43	0.1905
2	Gender†	0, 1	118, 67	0.0316
3	Age	40–92	62	0.0418
4	Medical center‡	0, 1	140, 45	0.1983
5	Volume of soft palate	1.5–12	4.56	−0.1184
6	Mean dose to soft palate (Gy)	0–73	25.96	0.4123
7	Volume of contralateral parotid gland	9–58	25.10	−0.1360
8	Mean dose to contralateral parotid glands (Gy)	0–69	26.37	0.4935
9	Volume of ipsilateral parotid gland	9–56	26.31	−0.1356
10	Mean dose to ipsilateral parotid gland(Gy)	0–71	35.16	0.4634
11	Volume of contralateral sublingual gland	0.06–2.5	0.39	−0.0571
12	Mean dose to contralateral sublingual gland (Gy)	0–73	14.53	0.3070
13	Volume of ipsilateral sublingual gland	0.06–2.4	0.35	−0.1355
14	Mean dose to ipsilateral sublingual gland (Gy)	0–73	16.31	0.2574
15	Volume of contralateral submandibular gland	3.5–17	8.65	−0.1228
16	Mean dose to contralateral submandibular gland (Gy)	0.05–76	52.70	0.4532
17	Volume of ipsilateral submandibular gland	0.6–20	8.71	−0.0886
18	Mean dose to ipsilateral submandibular gland (Gy)	0.06–76	54.71	0.4362
19	Volume of the lower lip	0.5–5.5	1.98	−0.0013
20	Mean dose to the lower lip(Gy)	0–71	0.79	0.1662
21	Baseline xerostomia score(s)	1, 2, 3, 4	106, 53, 19, 7	0.3578

* Chemotherapy = 0, No; 1, Yes.
† Gender = 0, Male; 1, Female.
‡ Medical center = 0, University Medical Center Groningen; 1, Free University Medical Center Amsterdam.

BIG DATA

Lessons From Large-Scale Collection of Patient-Reported Outcomes: Implications for Big Data Aggregation and Analytics

Jeff A. Sloan, PhD,* Michele Halyard, MD,† Issam El Naqa, PhD,† and Charles Mayo, PhD†

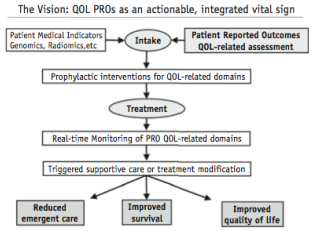


Fig. 1. Vision for use of patient-reported outcomes (PROs) as part of clinical decision frameworks. *Abbreviation:* QOL = quality of life.

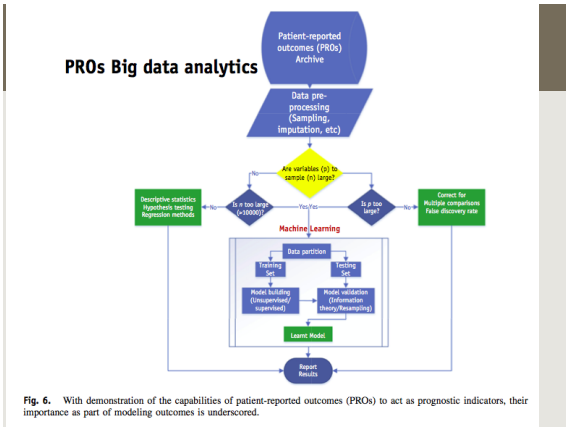


Fig. 6. With demonstration of the capabilities of patient-reported outcomes (PROs) to act as prognostic indicators, their importance as part of modeling outcomes is underscored.

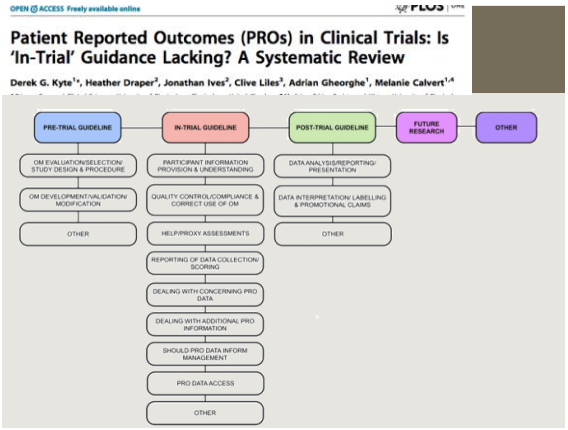


Table 1
Radiotherapy research data types within their common IT systems.

Information type	Data examples	IT system
Baseline clinical data	Demographics (including co-morbidity and family history), TNM-stage, date of diagnosis, histopathology	HIS, TDS
Diagnostic imaging data	Diagnostic CT, MR and PET imaging	PACS
Radiotherapy treatment planning data	Delineation/structure sets, planning-CT, dose matrix, beam set-up, prescribed dose and fractions	PACS, RIS
Radiotherapy treatment delivery data	Cone beam CTs, orthogonal EPID imaging, delivered fractions	PACS, RIS
Non-radiotherapy treatment data	Surgery, chemotherapy	HIS, TDS
Outcome data	Survival, local control, distant failure, toxicity (including patient reported outcomes), quality of life	EDC, TDS
Follow-up imaging data	Follow-up CT, MR and PET imaging	PACS
Biological data	Sample storage, shipping, tracing and lab results	LIMS
Additional study conduct data	Study design, protocol, eligibility criteria	EDC, CTMS

Data collection

Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial

Erik Roelofs [a,*,1], Lucas Persoon [a,1], Sebastiaan Nijsten [a], Wolfgang Wiessler [b], André Dekker [a,1], Philippe Lambin [a,1]

[a] Department of Radiation Oncology (MAASTRO Clinic), Maastricht University Medical Centre (MUMC+), The Netherlands; [b] Siemens Healthcare, Malvern, PA, USA

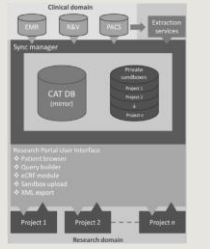


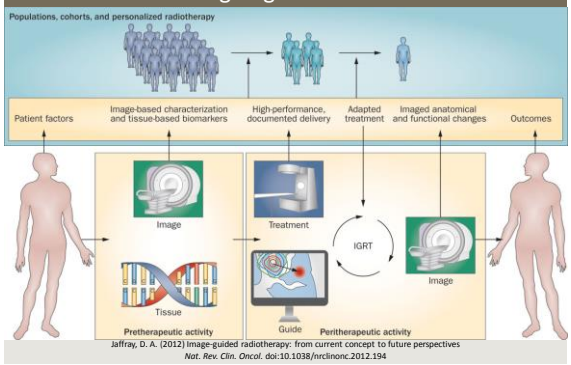
Fig. 1. Schematic overview of the CAT data warehouse/research portal. The system synchronizes data from clinical data sources and custom services. It is also capable of collecting data for trials and data collected for other research purposes. For data export, several modules exist in the system and are easily accessible by web-technology (i.e. the patient browser, query builder and an electronic case report form XML export).

Table 1
Parameters collected for the NSCLC and rectal cancer groups. The last columns show which data were looked up where and from which source the data were recalculated.

Parameter	NSCLC	Rectum	Source		Action
			Manual	Automatic	
Gender	√	√	Chart	EMR	Looked up
WHO score	√	√	Chart	EMR	
TNM staging	√	√	Chart	EMR	
Chemo therapy	√	√	Chart	EMR	
Number of positive lymph nodes	√	√	Chart	EMR	
Tumour PA	√	√	Chart	EMR	
pCR		√	Chart	EMR	
Survival	√	√	Chart	EMR	
Total delivered dose	√	√	R&V	R&V	
Overall treatment time	√	√	R&V	R&V	
GTV volume	√	√	XiO	PACS	Recalculated
V_5	Lungs[a]		XiO	PACS	
V_{20}	Lungs		XiO	PACS	
V_{40}		Bladder	XiO	PACS	
MLD	√[b]		XiO	PACS	
SUV Max		Tumour	TrueD	PACS	
SUV Mean		Tumour	TrueD	PACS	

[a] V_5 and V_{20} data for the lungs were calculated with both lungs minus the PTV.
[b] MLD data for the lungs were calculated with both lungs minus the GTV.

Integration of imaging information in designing treatments



Jaffray, D. A. (2012) Image-guided radiotherapy: from current concept to future perspectives *Nat. Rev. Clin. Oncol.* doi:10.1038/nrclinonc.2012.194

Radiomics/radiogenomics magnifies this challenge/opportunity

- Radiomics is "extraction and analysis of large amounts of advanced quantitative imaging features with high throughput from medical images obtained with computed tomography, positron emission tomography or magnetic resonance imaging." (Kumar, 2012)
- Radiogenomics
  - In imaging community refers to linkage of imaging features w genomics
  - In radiation oncology community refers to genomic correlates of radiation response

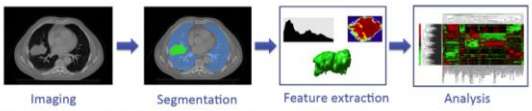


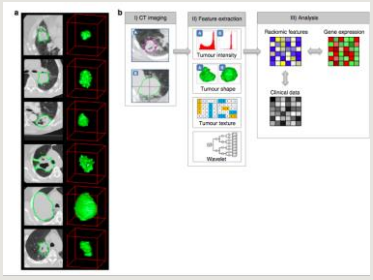
Fig. 4. The Radiomics workflow. On the medical images, segmentation is performed to define the tumour region. From this region the features are extracted, e.g. features based on tumour intensity, texture and shape. Finally, these features are used for analysis, e.g. the features are assessed for their prognostic power, or linked with stage, or gene expression.

Exciting time!!

Received 25 Nov 2013 | Accepted 29 Apr 2014 | Published 3 Jun 2014 | Updated 7 Aug 2014

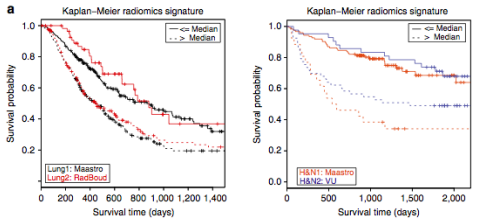
Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach

Hugo J.W.L. Aerts[1,2,3,4], Emmanuel Rios Velazquez[2,3,4], Ralph T.H. Leijenaar[1], Chintan Parmar[1,2], Patrick Grossmann[2], Sara Cavalho[1], Johan Bussink[5], René Monshouwer[5], Benjamin Haibe-Kains[6], Derek Rietveld[7], Frank Hoebers[7], Michelle M. Rietbergen[8], C. René Leemans[8], Andre Dekker[7], John Quackenbush[4], Robert J. Gillies[9] & Philippe Lambin[1]



Wow!

NATURE COMMUNICATIONS | DOI: 10.1038/ncomms5006



Lung Texture in Serial Thoracic Computed Tomography Scans: Correlation of Radiomics-based Features With Radiation Therapy Dose and Radiation Pneumonitis Development

Alexandra Cunliffe, PhD,* Samuel G. Armato III, PhD,* Richard Castillo, PhD,† Ngoc Pham, MD,‡ Thomas Guerrero, MD, PhD,‡ and Hania A. Al-Hallaq, PhD†

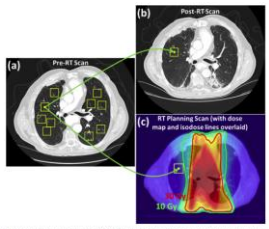


Fig. 1. Process for mapping the center of each ROI (yellow box) from the pre-RT CT scan (a) to the post-RT CT scan (b) and the RT planning CT scan with associated dose map (c). CT scans. CT = computed tomography; ROI = region of interest; RT = radiation therapy. A color version of this figure is available at www.redjournal.org.

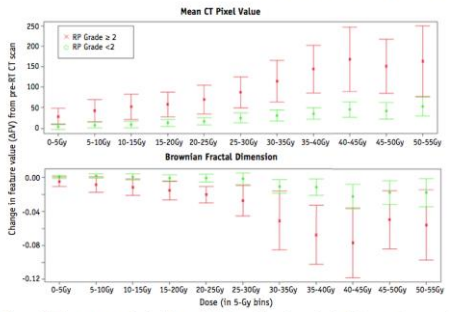


Fig. 2. Plots of ΔFV versus mean physical dose with 95% confidence intervals for 2 features. Average values were calculated on a per-patient basis in 5-Gy dose bins.

Table 4 Comparison of single-feature performance to distinguish patients with RP compared with a classifier composed of each 2- or 3-feature combination

No. of features/dose-dependent measurement	Mean AUC (range)			
	Low dose	Medium dose	High dose	Fitted slope
1 Feature	0.64 (0.49-0.72)	0.68 (0.49-0.77)	0.71 (0.52-0.78)	0.71 (0.50-0.78)
2 Features	0.66 (0.59-0.74)	0.71 (0.59-0.78)	0.73 (0.59-0.78)	0.74 (0.59-0.84)
3 Features	0.66 (0.59-0.75)	0.72 (0.59-0.79)	0.72 (0.60-0.79)	0.75 (0.61-0.83)

Abbreviations: AUC = area under the receiver operating characteristic curve; RP = radiation pneumonitis.

Conclusions

This study demonstrated that quantitative measurement of dose-dependent texture changes between pre- and post-RT CT scans can differentiate between patients with clinical (grade ≥2) RP and those patients without RP. Twelve intensity- and texture-based features demonstrated significantly increased changes for patients with RP. In general, individual features could be used to discriminate between patients with and without RP with moderate performance. When multiple features were combined in a classifier, AUC increased significantly (AUC values from 0.59 to 0.84). This study demonstrates the potential ability of radiomics to provide a quantitative, individualized approach to measure patient lung tissue reaction to RT and assess radiation pneumonitis.

Radiotherapy and Oncology 118 (2015) 462–466

Contents lists available at ScienceDirect

Radiotherapy and Oncology

journal homepage: www.thegreenjournal.com

Experimental radiobiology

Is there a causal relationship between genetic changes and radiomics-based image features? An *in vivo* preclinical experiment with doxycycline inducible GADD34 tumor cells

Kranthi Marella Panth[a,]*, Ralph T.H. Leijenaar[1], Sara Carvalho, Natasja G. Lieuwes, Ala Yaromina, Ludwig Dubois, Philippe Lambin

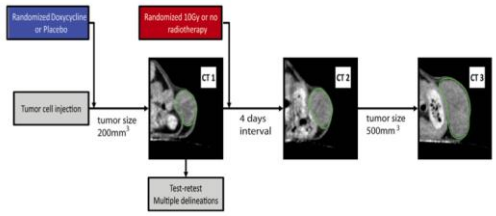


Fig. 1. Graphical representation of experimental work flow.

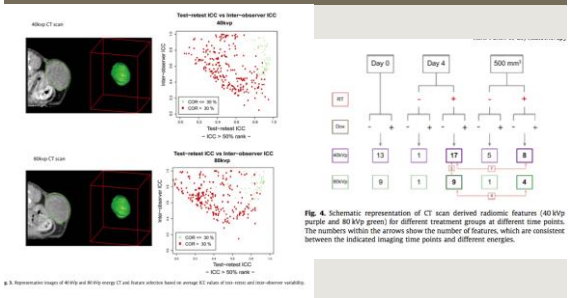


Fig. 4. Schematic representation of CT scan derived radiomic features (40 kVp purple and 80 kVp green) for different treatment groups at different time points. The numbers within the arrows show the number of features, which are consistent between the indicated imaging time points and different energies.

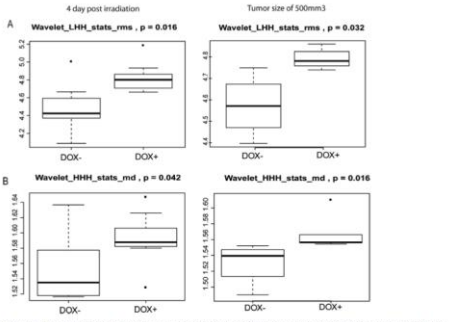


Fig. 5. Gene induction (dox+) results in significantly changed features compared to no gene induction (dox−) upon combination with radiotherapy 4 days after RT and at 500 mm³. Features are extracted from 40 kVp (A) or 80 kVp (B) CT scans.

However, when GADD34 was induced and radiation therapy was given, there was additional growth delay due to a decreased hypoxic fraction at the time of irradiation, as previously described [12]. Thus the difference in tumor heterogeneity between gene-induced and non-induced tumors post RT was reflected in the imaging features likely as a result of a phenotypic change. Interestingly, the radiomics image features that were found to be significantly different between both groups shortly after irradiation were also observed at larger tumor volumes. This phenomenon was observed independent of the CT image acquisition energy level, although the observed features were different between both energies tested. Remarkably, the feature value for slow-growing tumors (gene-induced) was higher than for faster-growing tumors (no gene-induced group) upon combination with radiotherapy.

Conclusion

We have shown in *in vivo* preclinical models that radiomics is able to quantify the early effects of radiation treatment and genetic changes in tumors with similar volumes, and identify differences that are not visible to the human eye.

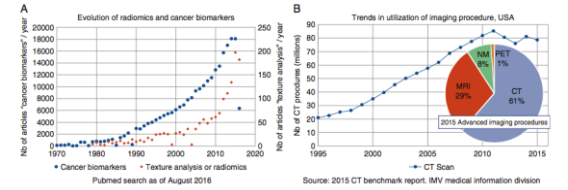
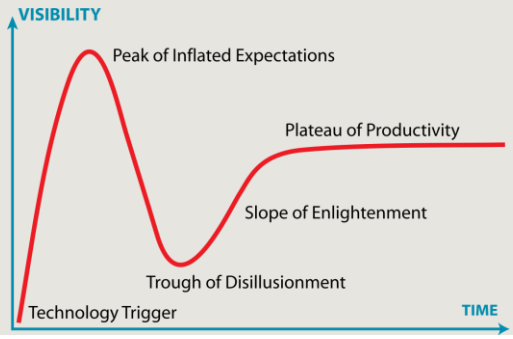
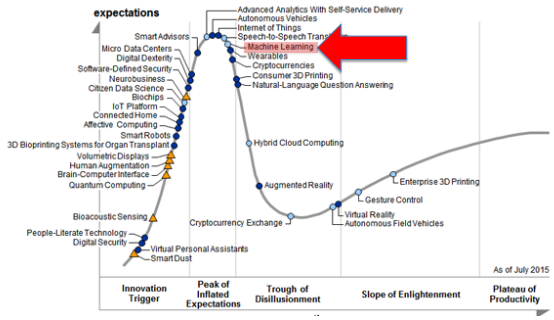


Figure 2. (A) Comparison of publications on radiomics and molecular biomarkers. (B) Trends in imaging utilization in the USA. CT, Computed tomography scan; MRI, Magnetic resonance imaging; NM, Nuclear medicine; PET, Positron emission tomography.

But let's not get carried away in a "hype cycle"…



Watch out for the trough!!

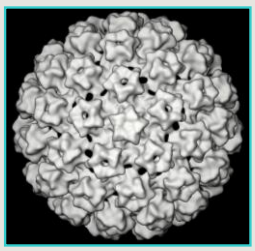


Example: Oropharynx Machine Learning/Radiomics Challenge

- ~20,000 US cases
- Oropharynx cancer (OPC) is epidemic due to human papilloma virus (HPV) infection
- More HPV-associated head and neck cancers than cervical cancers now

	ESTIMATED NEW CASES			ESTIMATED DEATHS		
	BOTH SEXES	MALE	FEMALE	BOTH SEXES	MALE	FEMALE
All Sites	**1,529,560**	**789,620**	**739,940**	**569,490**	**299,200**	**270,290**
Oral cavity & pharynx	**36,540**	**25,420**	**11,120**	**7,880**	**5,430**	**2,450**
Tongue	10,990	7,690	3,300	1,990	1,300	690
Mouth	10,840	6,430	4,410	1,830	1,140	690
Pharynx	12,660	9,880	2,780	2,410	1,730	680
Other oral cavity	2,050	1,420	630	1,650	1,260	390

Human Papillomavirus (HPV)

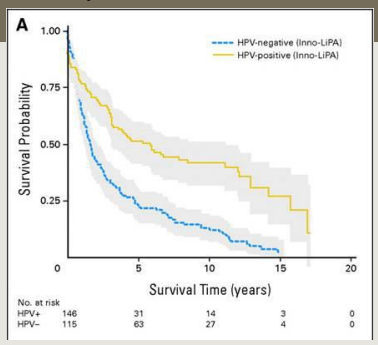


- **DNA virus**
- **>100 different sub-types**
- **Infects skin and mucosa**
- **Asymptomatic**
- **Benign growths – warts**
- **Oncogenic (cancer causing) types are mostly 16 and 18**

Slide courtesy Adam Garden

Follow-up for overall survival

271 samples from SEER



Chaturvedi A K et al. JCO 2011;29:4294-4301 JOURNAL OF CLINICAL ONCOLOGY

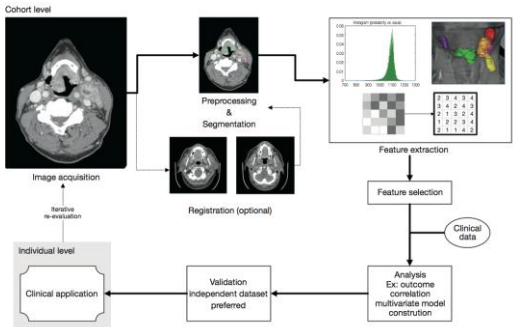


Figure 1 The "radiomics workflow" involves a series of iterative steps for reproducible and consistent extraction of imaging data. These steps include image acquisition, tumor segmentation, feature extraction, and feature selection. The selected features can then be analyzed for outcome correlation and potential incorporation into predictive models. Additionally, validations should be done against completely independent large datasets, preferably from other institutions.

Transl Cancer Res 2016;5(4):371-382

Radiomics in head and neck cancers: Classification

Table 1 Studies on radiomics for segmentation and classification

Authors (study)	Publication date	Modality	# of patients	Anatomic site, if specified	Analyzed endpoint
Raja *et al.* (23)	Sep 2012	CT	21	Oral cavity	Tumor grade classification
Buch *et al.* (17)	Jul 2015	CT	40	Oropharynx	HPV status
Fujita *et al.* (24)	Jan 2016	CT	46	Oropharynx (25); larynx (17); hypopharynx (5)	HPV status
Yu *et al.* (26)	Mar 2009	FDG-PET/CT	20	–	Normal vs. abnormal tissue classification
Yu *et al.* (27)	Oct 2009	FDG-PET/CT	10	Oropharynx and nasopharynx	Normal vs. abnormal tissue classification
Vallieres *et al.* (25)	Oct 2013	FDG-PET	67	–	HPV status; loco-regional failure; distant metastasis
Fruehwald-Pallamar *et al.* (28)	Nov 2013	MRI	38	Parotid	Benign vs. malignant status; tumor type differentiation
Yang *et al.* (29)	Dec 2014	MRI	15	Parotid	Parotid vs. surrounding tissue differentiation
Brown *et al.* (15)	May 2015	DW-MRI	26 (training) 18 (validation)	Thyroid	Thyroid nodule classification
Jansen *et al.* (21)	Jan 2016	DCE-MRI	19	Oropharynx	Local control; local failure
Park *et al.* (30)	Feb 2016	MRI	27	Oropharynx	Tumor type differentiation
Fruehwald-Pallamar *et al.* (31)	Feb 2016	MRI	100	–	Benign vs. malignant status

CT, computed tomography; HPV, human papillomavirus; FDG-PET, fludeoxyglucose-positron emission tomography; MRI, magnetic resonance imaging; DW, diffusion-weighted; DCE, dynamic contrast-enhanced.

Transl Cancer Res 2016;5(4):371-382

Head and neck radiomics: Outcomes

Table 2 Studies on radiomics for prognostic and predictive biomarkers

Authors (study)	Publication date	Modality	# of patients	Anatomic site, if specified	Analyzed endpoint
Zhang *et al.* (22)	Dec 2013	CT	72	Oral cavity (28); larynx (21); hypopharynx (14); oral cavity (8)	Overall survival
Aerts *et al.* (34)	Jun 2014	CT	474 (training); 545 (validation)	Lung or head and neck	Median survival
Parmar *et al.* (35)	Jun 2015	CT	878	Lung or head and neck	Survival; tumor stage; HPV status
Parmar *et al.* (36)	Dec 2015	CT	101 (training); 95 (validation)	–	Overall survival
Leijenaar *et al.* (37)	Aug 2015	CT	542	Oropharynx	Median survival
El Naqa *et al.* (20)	Jun 2009	FDG-PET	9	–	Overall survival
Dang *et al.* (38)	Jan 2015	MRI	16	Oropharynx	p53 status

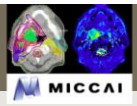
CT, computed tomography; HPV, human papillomavirus; FDG-PET, fludeoxyglucose-positron emission tomography; MRI, magnetic resonance imaging.

Our challenges…

- If image features in the local region can predict local phenotypic alteration (per Panth et al.), can we predict local oncologic outcomes (i.e. local control in the priamry tumor)?
- If the major driver of local control probability genomically is known, can we identify genetic correlates of local control risk (i.e. HPV status)?

Unmet needs, unanswered questions

- Can we crowdsource innovation?
- Can we draw on “non-insider” knowledge?
- Can we have some fun?

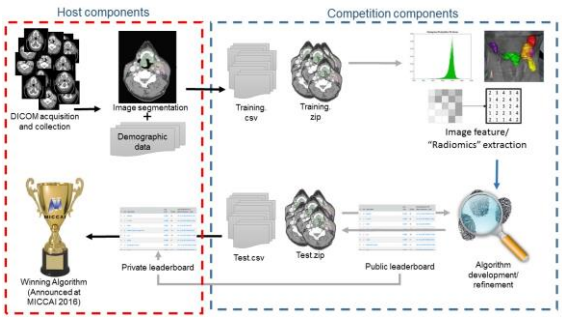


2016 OPC CPM challenges

Determine from CT data whether a tumor will be controlled by definitive radiation therapy.

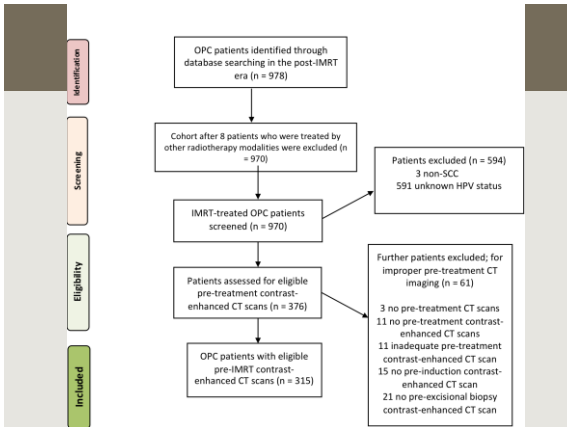
Predict from CT data the HPV phenotype of oropharynx tumors; compare to ground-truth results previously obtained by p16 or HPV testing.

Kaggle/MICCAI challenges



Rules of the game

- Binary classification
  - HPV status (1/0)
  - Local control of tumor in the primary site (1/0)
- Can leverage demographic data
- Half dataset with known HPV/local control available for training
  - Other half randomly split into "public" validation set (visible to end users) and "private" validation set
  - Could submit and see performance on "public" set to avoid overfitting
  - Winner declared on "private" dataset

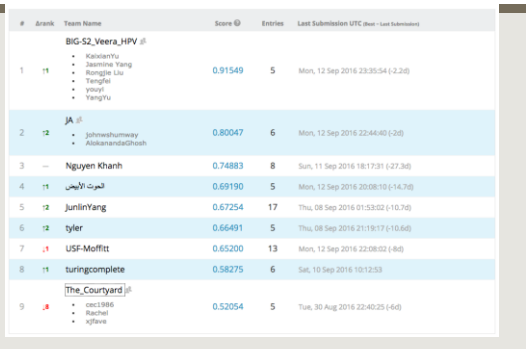


Additionally, 1st, 2nd, and 3rd place teams will receive a commemorative trophy, ideally of limited cash value and likely with hideous design, allowing gentle flaunting of their definitive victory over anguished peers and colleagues who failed to place in the challenge. An example is shown below.

Results

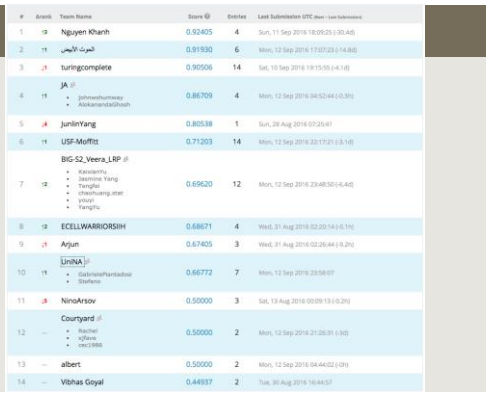
Local Control Challenge	HPV Challenge
14 teams	9 teams
23 players	17 players
78 entries	70 entries

Private Leaderboard - Oropharynx Cancer (OPC) Radiomics Challenge :: Human Papilloma Virus (HPV) Status Prediction



#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best - Last Submission)
1	↑1	BIG-S2_Veera_HPV • KaixianYu • Jasmine Yang • Rongjie Liu • Tengfei • youyi • YangYu	0.91549	5	Mon, 12 Sep 2016 23:35:54 (-2.2d)
2	↑2	JA • Johnwshumway • AlokanandaGhosh	0.80047	6	Mon, 12 Sep 2016 22:44:40 (-2d)
3	–	Nguyen Khanh	0.74883	8	Sun, 11 Sep 2016 18:17:31 (-27.3d)
4	↑1	العوث الأبيض	0.69190	5	Mon, 12 Sep 2016 20:08:10 (-14.7d)
5	↑2	JunlinYang	0.67254	17	Thu, 08 Sep 2016 01:53:02 (-10.7d)
6	↑2	tyler	0.66491	5	Thu, 08 Sep 2016 21:19:17 (-10.6d)
7	↓1	USF-Moffitt	0.65200	13	Mon, 12 Sep 2016 22:08:02 (-8d)
8	↑1	turingcomplete	0.58275	6	Sat, 10 Sep 2016 10:12:53
9	↓8	The_Courtyard • cec1986 • Rachel • kjflave	0.52054	5	Tue, 30 Aug 2016 22:40:25 (-6d)

Oropharynx Cancer (OPC) Radiomics Challenge :: Local Recurrence Prediction



#	Δrank	Team Name	Score	Entries	Last Submission UTC (Best - Last Submission)
1	↑2	Nguyen Khanh	0.92405	4	Sun, 11 Sep 2016 18:09:25 (-30.4d)
2	↑1	العوث الأبيض	0.91930	6	Mon, 12 Sep 2016 17:07:23 (-14.8d)
3	↓1	turingcomplete	0.90506	14	Sat, 10 Sep 2016 19:15:55 (-4.1d)
4	↑1	JA • Johnwshumway • AlokanandaGhosh	0.86709	4	Mon, 12 Sep 2016 04:52:44 (-0.3h)
5	↓4	JunlinYang	0.80538	1	Sun, 28 Aug 2016 07:25:41
6	↑1	USF-Moffitt	0.71203	14	Mon, 12 Sep 2016 22:17:21 (-3.1d)
7	↑2	BIG-S2_Veera_LRP • KaixianYu • Jasmine Yang • Tengfei • chaohuang.stat • youyi • YangYu	0.69620	12	Mon, 12 Sep 2016 23:48:50 (-6.4d)
8	↑2	ECELLWARRIORSIIH	0.68671	4	Wed, 31 Aug 2016 02:20:14 (-5.1h)
9	↓1	Arjun	0.67405	3	Wed, 31 Aug 2016 02:26:44 (-9.2h)
10	↑1	UniNA • GabrielePantaleo • Stefano	0.66772	7	Mon, 12 Sep 2016 23:58:07
11	↓3	NinoArsov	0.50000	3	Sat, 13 Aug 2016 00:09:13 (-0.2h)
12	–	Courtyard • Rachel • kjflave • cec1986	0.50000	2	Mon, 12 Sep 2016 21:26:31 (-3d)
13	–	albert	0.50000	2	Mon, 12 Sep 2016 04:44:02 (-0h)
14	–	Vibhas Goyal	0.44937	2	Tue, 30 Aug 2016 16:44:57

Congrats to our winners!!

- **1st Place** and recipient of **a proffered manuscript acceptance (post-editorial review), with waiver of fees** in the well-renowned international, open access ESTRO-sponsored journal; Clinical and Translational Radiation Oncology (ctRO):

Nguyen Khanh ← This college kid beat everyone else by *NOT USING RADIOMICS*

Congrats Nguyen!! But The best is yet to come. **You're invited to attend the 19th MICCAI international conference**, taking place in Athens, Greece during the period October 17-21, 2016 in person **(no travel support will be provided)** or via Skype to present your winning algorithm and be celebrated among your fellow data scientists.

- **1st Place** and recipient of **a proffered manuscript acceptance (post-editorial review), with waiver of fees** in the well-renowned international, open access ESTRO-sponsored journal; Clinical and Translational Radiation Oncology (ctRO):

BIG-S2_Veera_HPV

Congrats!! But The best is yet to come. **You're invited to attend the 19th MICCAI international conference**, taking place in Athens, Greece during the period October 17-21, 2016 in person **(no travel support will be provided)** or via Skype to present your winning algorithm and be celebrated among your fellow data scientists.

TRUTH: Good models are hard to find…so showing your work helps :)

Machine Learning methods for Quantitative Radiomic Biomarkers

Chintan Parmar[1,2,3,*], Patrick Grossmann[1,2,3], Johan Bussink[4], Philippe Lambin[1] & Hugo J. W. L. Aerts[1,2,3]

Classification method acronym	Classification method name	Feature selection method acronym	Feature selection method name
Nnet	Neural network	RELF	Relief
DT	Decision Tree	FSCR	Fisher score
BST	Boosting	GINI	Gini index
BY	Bayesian	CHSQ	Chi-square score
BAG	Bagging	JMI	Joint mutual information
RF	Random Forset	CIFE	Conditional infotrax feature extraction
MARS	Multi adaptive regression splines	DISR	Double input symmetric relevance
SVM	Support vector machines	MIM	Mutual information maximization
DA	Discriminant analysis	CMIM	Conditional mutual information maximization
NN	Neirest neighbour	ICAP	Interaction capping
GLM	Generalized linear models	TSCR	T-test score
PLSR	Partial least squares and principal componenet regression	MRMR	Minimum redundancy maximum relevance
—	—	MIFS	Mutual information feature selection
—	—	WLCX	Wilcoxon

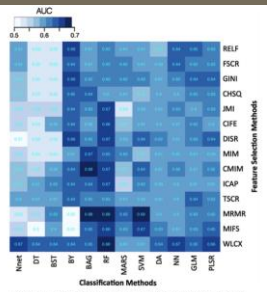


Figure 2. Heatmap depicting the predictive performance (AUC) of feature selection (in rows) and classification (in columns) methods. It can be observed that RF, BAG and BY classification methods and feature selection methods WLCX, MRMR and MIFS shows relatively high predictive performance in many cases.

Share your models/approaches!

- Github
  - Software repository
- Figshare
  - Data repository
- The Cancer Imaging Archive
  - Image repository
- Nature Scientific Data
  - Publish data/software descriptor on PubMed

Data Descriptor: Matched computed tomography segmentation and demographic data for oropharyngeal cancer radiomics challenges

SCIENTIFIC DATA | 4:170077 | DOI: 10.1038/sdata.2017.77

MICCAI/M.D. Anderson Cancer Center Head and Neck Quantitative Imaging Working Group*

Data category	Description
Patient ID	Numbers_given_randomly_to_the_patient_after_anonymizing_the_DICOM_PHI_tag_(0010,0020)_Patient_ID)
HPV/p16 status	Human Papilloma Virus status, as assessed by HPV DNA in situ hybridization[11] and/or p16 protein expression via immunohistochemistry (IHC), with the results described as 1 (i.e., Positive) or 0 (i.e., Negative)
Gender	Patient's sex
Age at diagnosis	Patient's age in years at the time of diagnosis
Race	American Indian/Alaska Native, Asian, Black, Hispanic, White or NA (Not applicable)
Tumor laterality	Right, left, bilateral
Oropharynx subsite of origin	The subsite of the tumor within the oropharynx, i.e., base of tongue (BOT), tonsil/soft palate/pharyngeal wall/glossopharyngeal sulcus (GPS)/other (no single subsite of origin could be identified)
T category	The T category describes the original (primary) tumor, as regard its size and extent, per the American Joint Committee on Caner (AJCC) and Union for International Cancer Control (UICC) cancer staging system. It could be T1, T2, T3, T4. https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx
N category	The N category describes whether or not the cancer has reached nearby lymph nodes, per the AJCC and UICC cancer staging system. It can be N0, N1, N2a, N2b, N2c or N3. https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx
AJCC Stage	AJCC cancer stage. https://cancerstaging.org/references-tools/Pages/What-is-Cancer-Staging.aspx
Pathological grade	The grade of tumor differentiation. It is described as: I, II, III, IV, I-II, II-III or NA (Not assessable)
Smoking status at diagnosis	Never, current, or former
Smoking Pack-Years	An equivalent numerical value of lifetime tobacco exposure. A pack year is defined as twenty cigarettes smoked every day for one year. (NA: Not Assessable) http://smokingpackyears.com/

Table 2. **Supplementary Information about the data provided for both challenges.**

Be FAIR, use the FORCE!

FORCE11
The Future of Research Communications and e-Scholarship

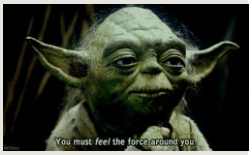
https://www.force11.org/node/6062

3. FAIR FOR MACHINES AS WELL A PEOPLE

In eScience, two clearly separated substrates for knowledge discovery can be distinguished.

1. The actual data, which is as a rule beyond human intellectual capacity to analyse and
2. The 'Explicitome' (everything we already made explicit in text, databases and any other format to date).

- Data should be **Findable**
- Data should be **Accessible**
- Data should be **Interoperable**
- Data should be **Re-usable.**



The Future is FAIR

Comment: The FAIR Guiding Principles for scientific data management and stewardship

Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

So, we have some work to do…

- We need
  - Better data aggregation & quality
  - More transparent modeling
  - Shared datasets
  - Standards, standards, standards!
  - In short, more FAIR-ness ☺

But our real need is…

- You, the medical physics community…

Effective data integration is not easy or fast… It's just better!

We are on the edge of the possible… looking towards the vistas before us!

Thanks!

Email questions/comments:

- cdfuller@mdanderson.org