Artificial Intelligence in Medicine 🤖🏥

Dennis Bontempi

AIM
Artificial Intelligence in Medicine Program

HARVARD MEDICAL SCHOOL | Dana-Farber Cancer Institute | BWH BRIGHAM AND WOMEN'S HOSPITAL | MGH 1811 MASSACHUSETTS GENERAL HOSPITAL

Dennis Bontempi (PhD @ UM & Harvard's AIM)

Hugo Aerts, Director @ AIM

"The Radiomics guy"

HARVARD MEDICAL SCHOOL

AIM
Artificial Intelligence in Medicine Program

Maastricht University

Artificial Intelligence in Medicine:
(My Experience Transitioning) from Computer Science to Clinical Studies

Dennis Bontempi

May 27, 2021

Abstract

Over the last decade, there has been a resurgence of interest in artificial intelligence (AI) applications in medicine. Driven by the advent of deep-learning algorithms, AI is poised to become a transformational force in healthcare.

However, most of the deep learning literature published in computer science journals nowadays lacks the depth to be considered helpful to real applications in the medical field. Researchers make dramatic claims that are often not tested in a clinical setting, therefore rendering such publications more proof-of-concept than legitimate scientific studies. More and more experts are pushing back against the AI hype, pointing out that many alleged advances in the field are based on flimsy evidence. Artificial intelligence has become mired in a hype cycle which could lead to a second AI winter.

In this presentation, I will discuss the fundamental differences between AI research in the two aforementioned fields (i.e., computer science and medicine), providing a general overview of all the steps required to develop and validate AI pipelines in the context of medical studies. Furthermore, I will briefly discuss which tools could be borrowed by the computer science (and, more in general, the engineering) community from the medical field to potentially improve the quality of the research in these areas.


AIM/Introduction
Medicine VS Engineering
Research
Medicine TO Engineering
What does the group focus on
What's different
Some of our research
What translates

AIM/Introduction

What does the group focus on

Medicine VS Engineering

What's different

Research

Some of our research

Medicine TO Engineering

What translates

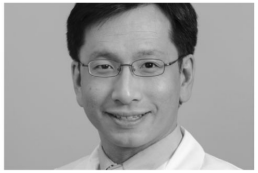
Artificial Intelligence in Medicine Program

Raymond Mak

Faculty Member

Danielle Bitterman

Faculty Member

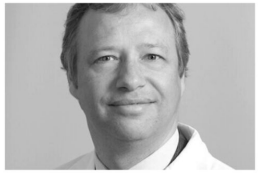


Michael Lu

Faculty Member

Benjamin Kann

Faculty Member



Udo Hoffmann

Faculty Member

Roman Zeleznik

Data Scientist

Ahmed Hosny

Data Scientist

Jakob Weiss

Research fellow

Elizabeth Huynh

Medical Physicist

Borek Foldyna

Affiliated Faculty

Zezhong Ye

Postdoctoral Fellow

Osbert Zalay

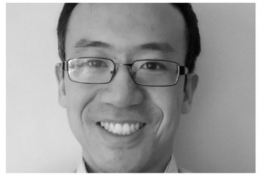
Research Fellow

Vineet Raghu

Postdoctoral Fellow

Christian Guthier

Physics Resident



Jack Qian

Resident

Márton Kolossváry

Postdoctoral Fellow

Subha Perni

Resident

Hannah Roberts

Resident

Tafadzwa Chaunzwa

Research Scholar

Yiwen Xu

Physics Resident

Zhongyi Zhang

Research Scholar

Ibrahim Hadžić

Research Scholar

Deborah Plana

Medical Student

Raymond Mak

Faculty Member

Danielle Bitterman

Faculty Member

Michael Lu

Faculty Member

Benjamin Kann

Faculty Member

Udo Hoffmann

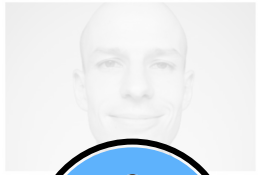
Faculty Member

Roman Zeleznik

Data Scientist

Ahmed Hosny

Data Scientist



Jak

Res

Eliz

Me

Borek Foldyna

Affiliated Faculty

Zezhong Ye

Postdoctoral Fellow

Osbert Zalay

Research Fellow

Vineet Raghu

Postdoctoral Fellow

Christian Guthier

Physics Resident

Jack Qian

Resident

Márton Kolossváry

Postdoctoral Fellow

Subha Perni

Resident

Hannah Roberts

Resident

Tafadzwa Chaunzwa

Research Scholar

Yiwen Xu

Physics Resident

Zhongyi Zhang

Research Scholar

Ibrahim Hadžić

Research Scholar

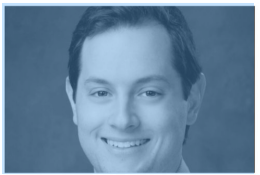
Deborah Plana

Medical Student

Raymond Mak
Faculty Member

Danielle Bitterman
Faculty Member

Michael Lu
Faculty Member



Benjamin Kann
Faculty Member

Udo Hoffmann
Faculty Member

Roman Zeleznik
Data Scientist

Ahmed Hosny
Data Scientist

Jakob Weiss
Research fellow

Elizabeth Huynh
Medical Physicist

Borek Foldyna
Affiliated Faculty

Zezhong Ye
Postdoctoral Fellow

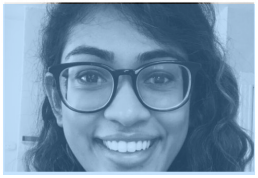
Osbert Zalay
Research Fellow

Vineet Raghu
Postdoctoral Fellow

Christian Guthier
Physics Resident

Jack Qian
Resident

Márton Kolossváry
Postdoctoral Fellow



Subha Perni
Resident

Hannah Roberts
Resident

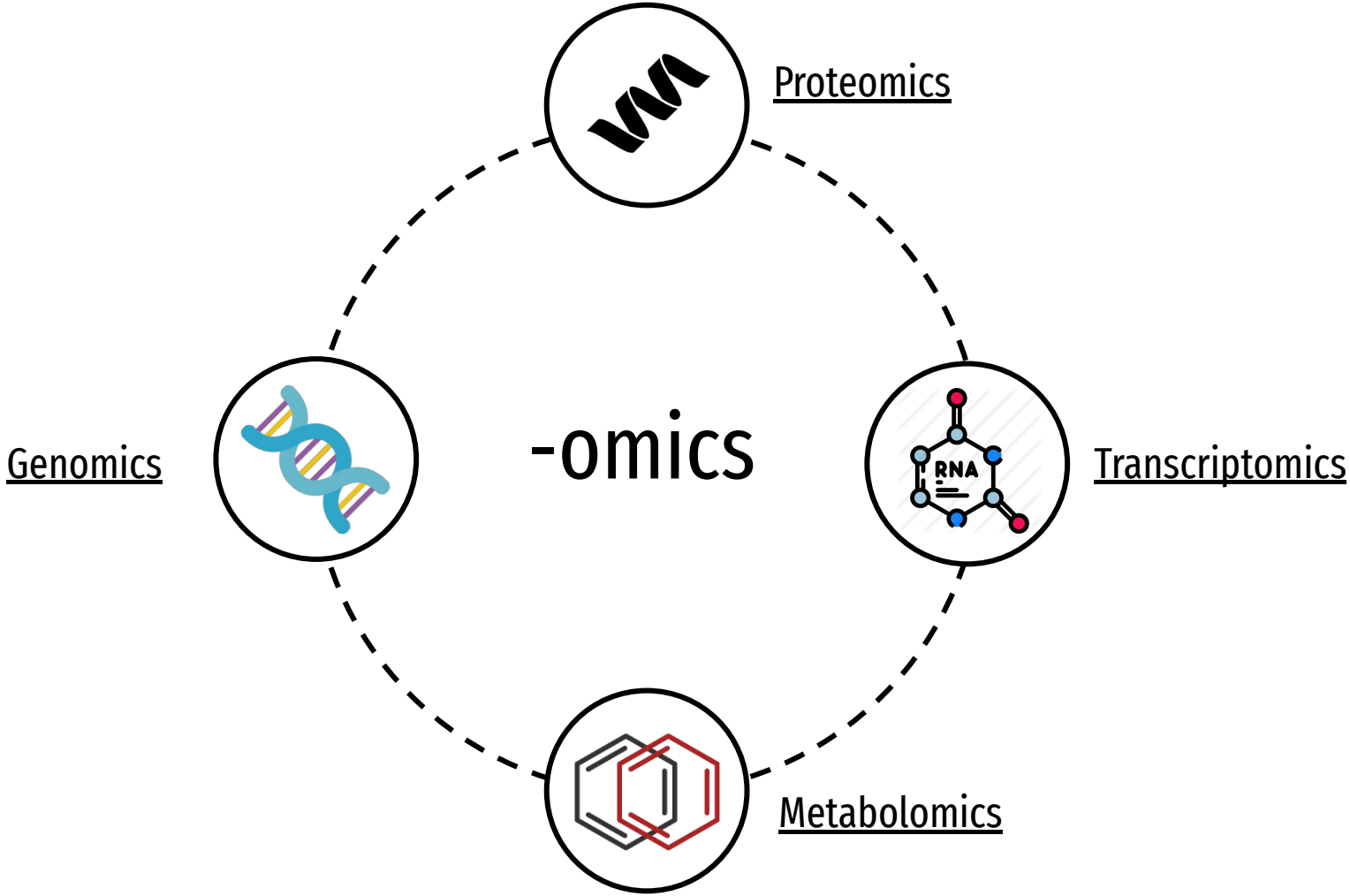
Tafadzwa Chaunzwa
Research Scholar

Yiwen Xu
Physics Resident

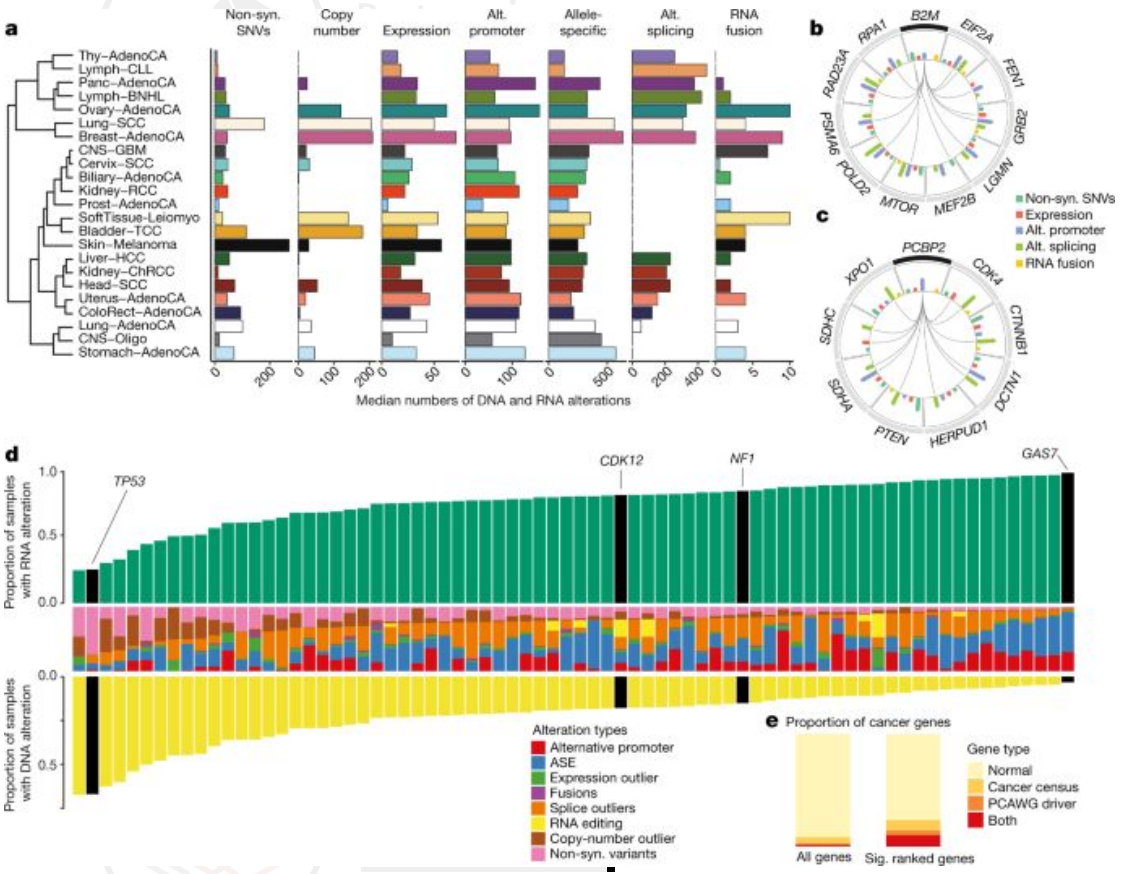
Zhongyi Zhang
Research Scholar

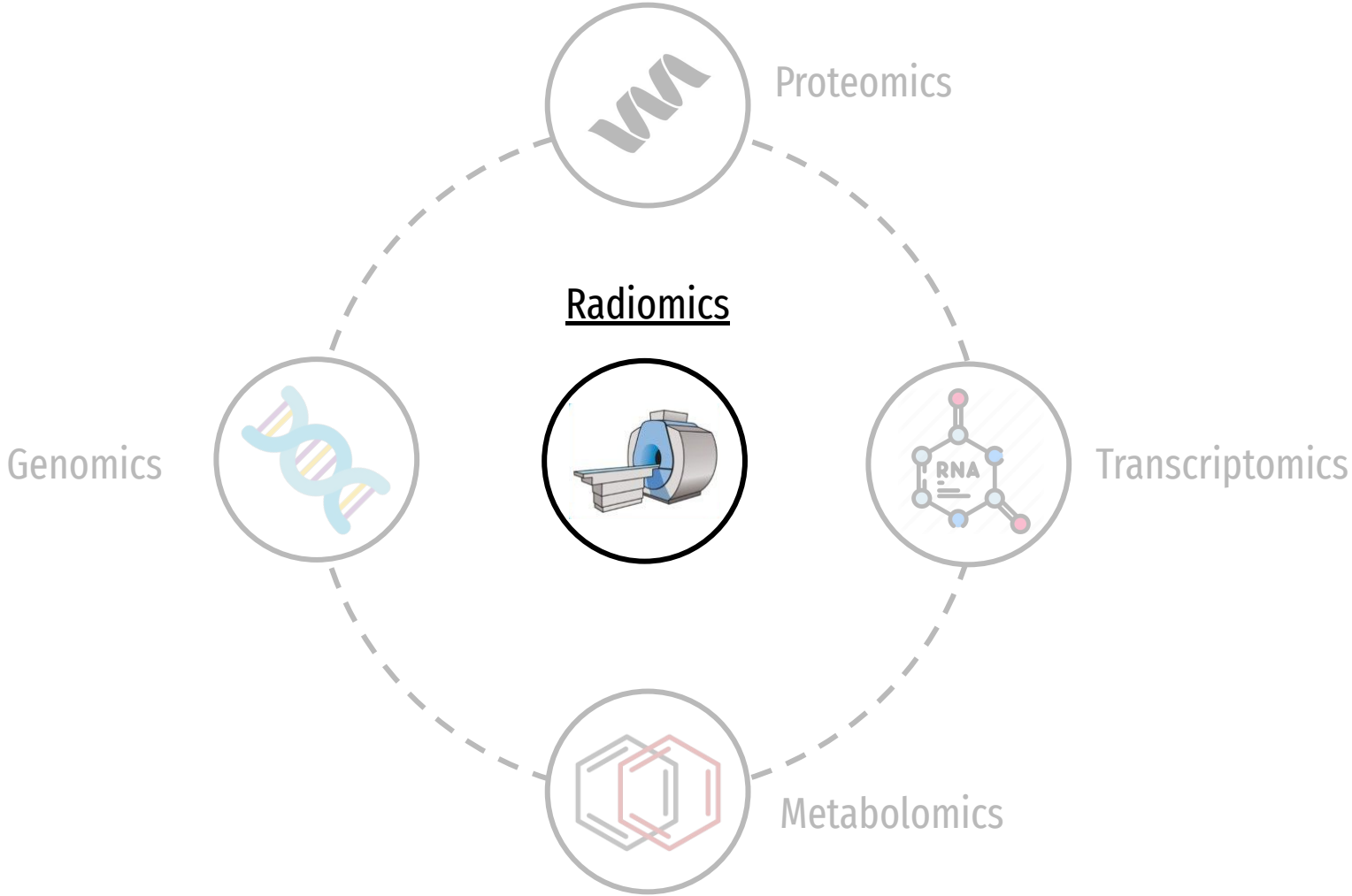
Ibrahim Hadžić
Research Scholar

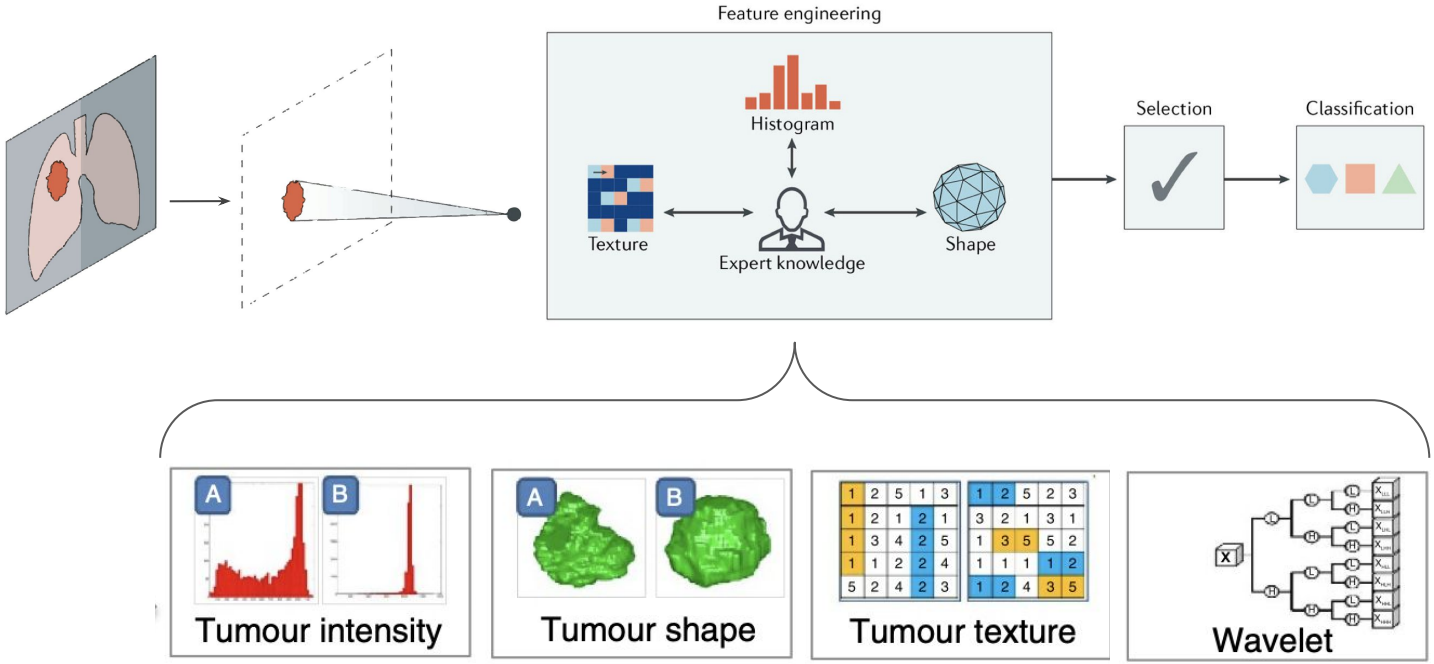
Deborah Plana
Medical Student


Proteomics
Genomics
-omics
RNA
Transcriptomics
Metabolomics

Genomics


a
Non-syn. SNVs
Copy number
Expression
Alt. promoter
Allele-specific
Alt. splicing
RNA fusion
Thy–AdenoCA
Lymph–CLL
Panc–AdenoCA
Lymph–BNHL
Ovary–AdenoCA
Lung–SCC
Breast–AdenoCA
CNS–GBM
Cervix–SCC
Biliary–AdenoCA
Kidney–RCC
Prost–AdenoCA
SoftTissue–Leiomyo
Bladder–TCC
Skin–Melanoma
Liver–HCC
Kidney–ChRCC
Head–SCC
Uterus–AdenoCA
ColoRect–AdenoCA
Lung–AdenoCA
CNS–Oligo
Stomach–AdenoCA
Median numbers of DNA and RNA alterations
b
B2M
EIF2A
FEN1
GRB2
LGMN
MEF2B
MTOR
POLD2
PSMA6
RAD23A
RPA1
c
PCBP2
CDK4
CTNNB1
DCTN1
HERPUD1
PTEN
SDHA
SDHC
XPO1
Non-syn. SNVs
Expression
Alt. promoter
Alt. splicing
RNA fusion
d
TP53
CDK12
NF1
GAS7
Proportion of samples with RNA alteration
Proportion of samples with DNA alteration
Alteration types
Alternative promoter
ASE
Expression outlier
Fusions
Splice outliers
RNA editing
Copy-number outlier
Non-syn. variants
e
Proportion of cancer genes
All genes
Sig. ranked genes
Gene type
Normal
Cancer census
PCAWG driver
Both


Proteomics
Radiomics
Genomics
RNA
Transcriptomics
Metabolomics



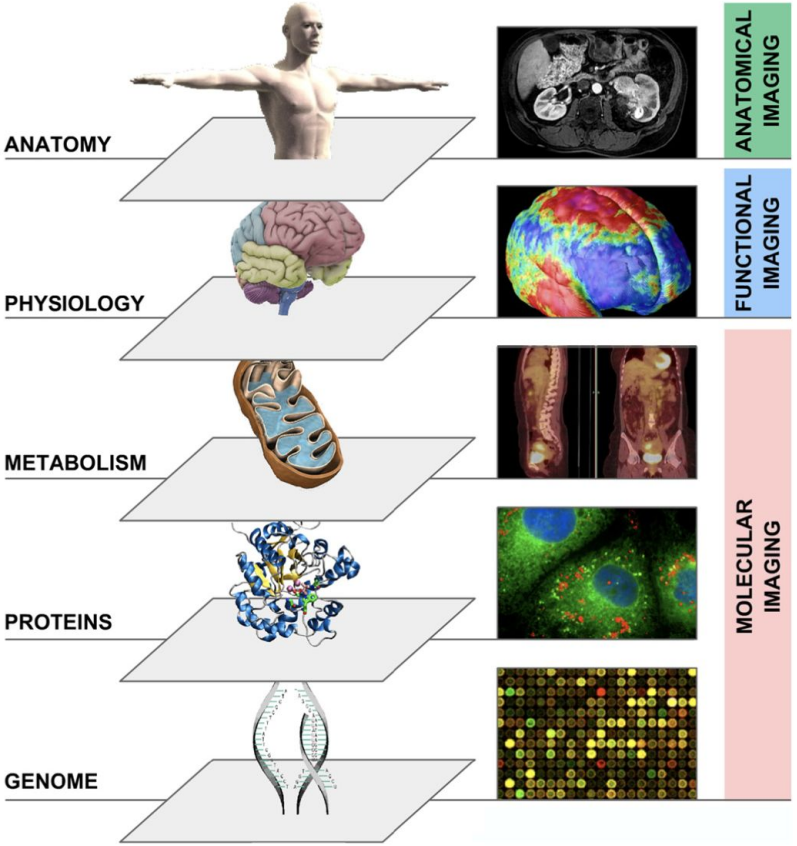
Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts

Artificial intelligence in radiology (NatRevCancer, 2018)

Qualitative → Quantitative

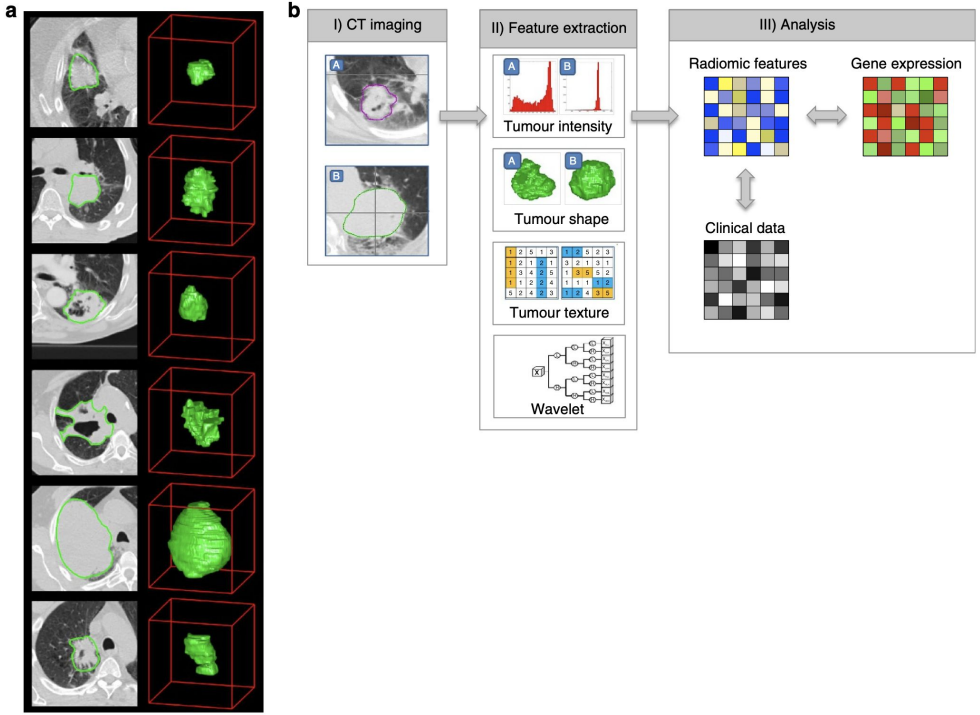
Ahmed Hosny, Chintan Parmar, John Quackenbush, Lawrence H. Schwartz, and Hugo J. W. L. Aerts

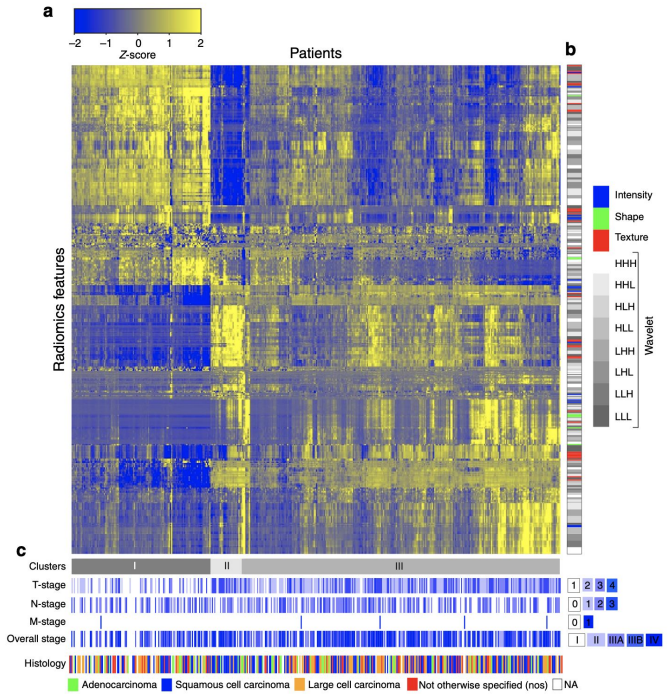
Artificial intelligence in radiology (NatRevCancer, 2018)



Philippe Lambin, Emmanuel Rios-Velazquez, Ralph Leijenaar, Sara Carvalho, et Al.

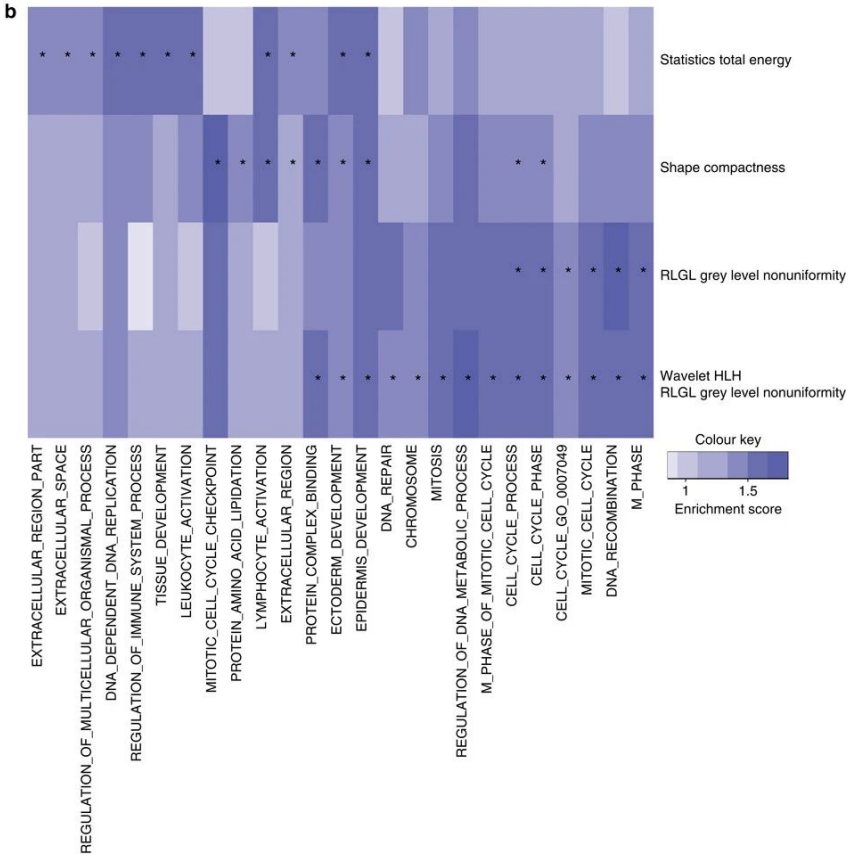
Radiomics: Extracting more information from medical image using advanced feature analysis (EU Journal of Cancer, 2012)

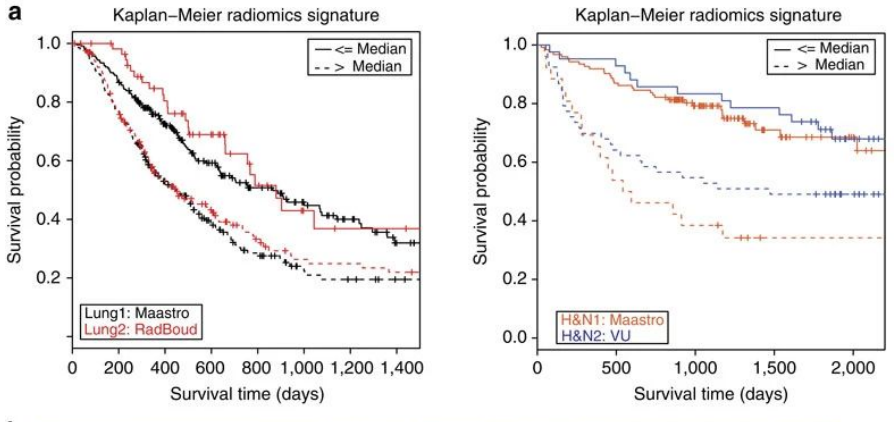




Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, et Al.

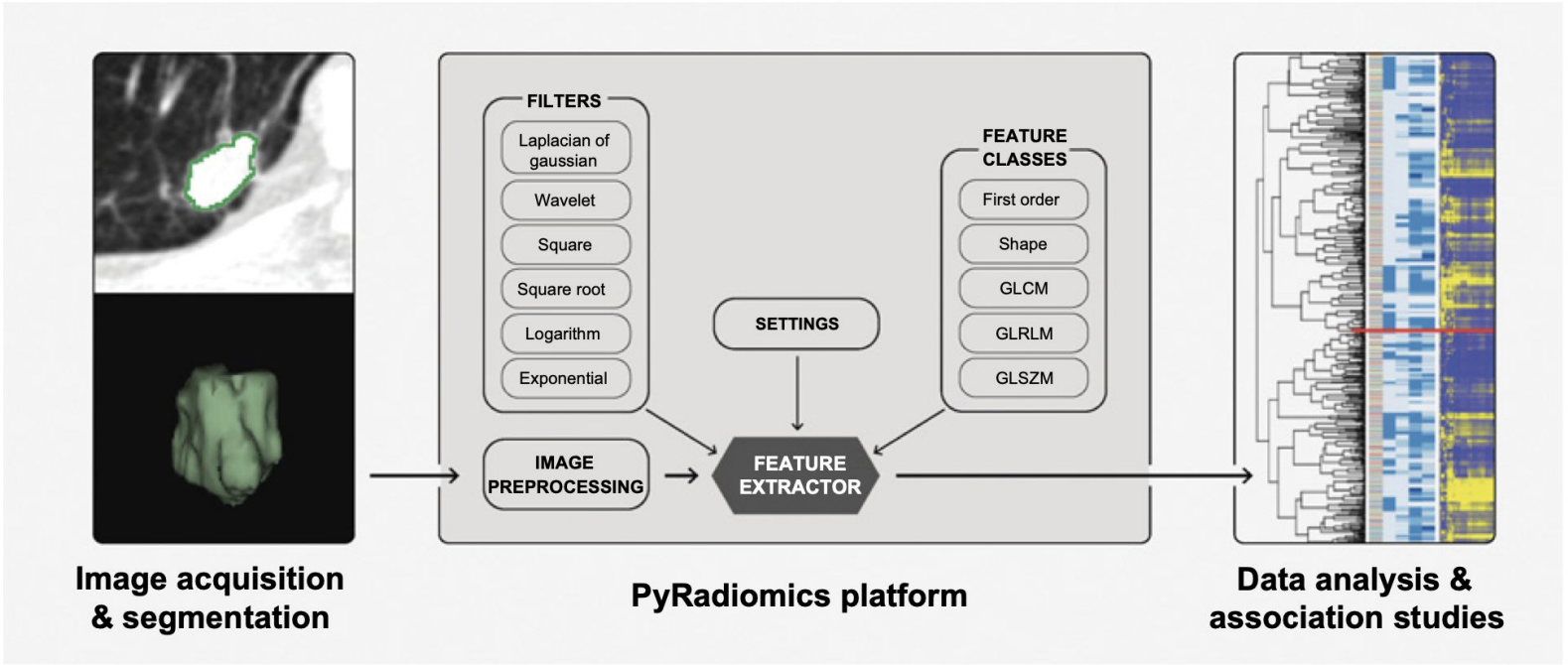
Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach (NatCom, 2014)





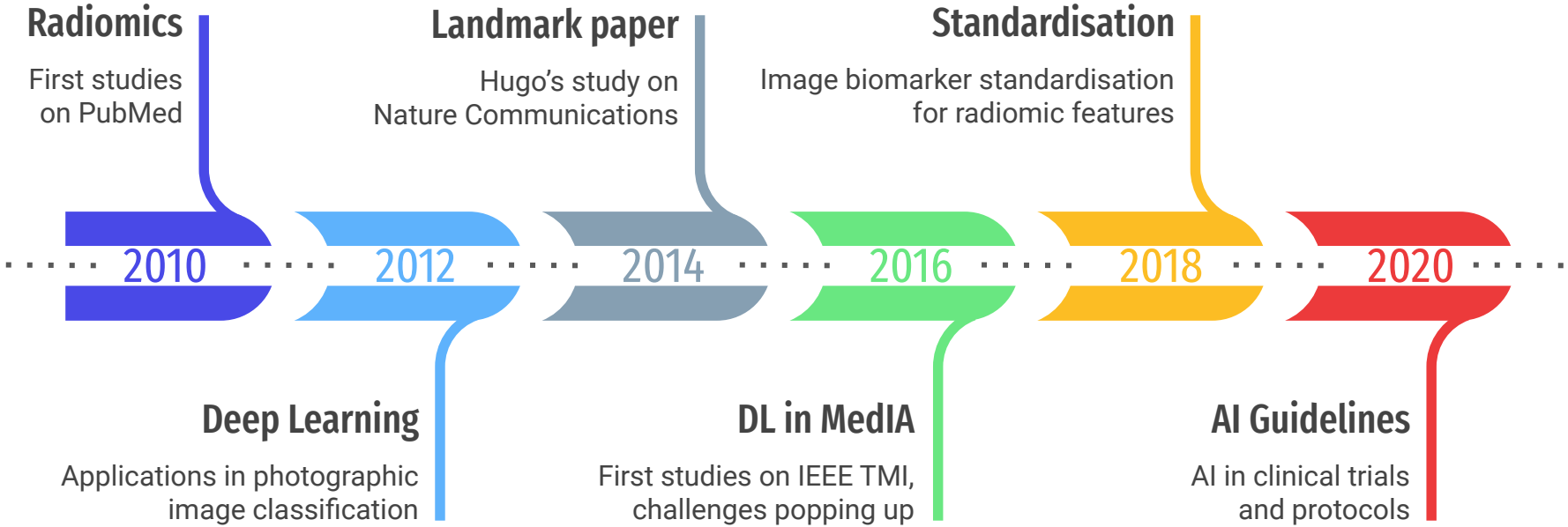
Hugo J. W. L. Aerts, Emmanuel Rios Velazquez, Ralph T. H. Leijenaar, Chintan Parmar, Patrick Grossmann, et Al.

Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach (NatCom, 2014)



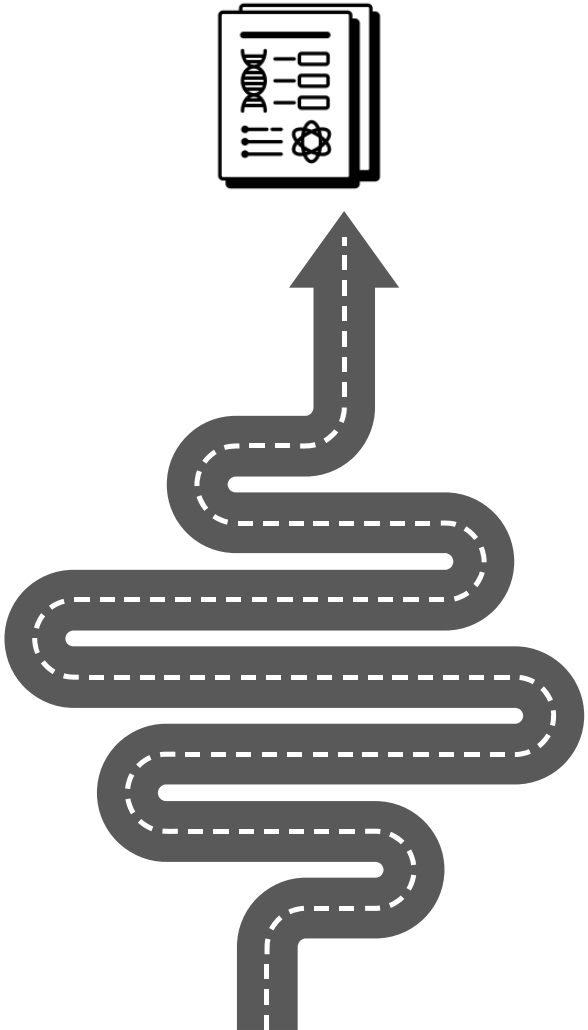
Joost J.M. van Griethuysen, Andriy Fedorov, Chintan Parmar, Ahmed Hosny, et Al.

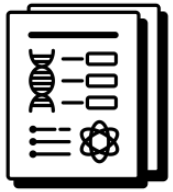
Computational Radiomics System to Decode the Radiographic Phenotype (Cancer Res, 2017)


Radiomics
First studies on PubMed
2010
Deep Learning
Applications in photographic image classification
2012
Landmark paper
Hugo's study on Nature Communications
2014
DL in MedIA
First studies on IEEE TMI, challenges popping up
2016
Standardisation
Image biomarker standardisation for radiomic features
2018
AI Guidelines
AI in clinical trials and protocols
2020


AIM/Introduction
Medicine VS Engineering
Research
Medicine TO Engineering
What does the group focus on
What's different
Some of our research
What translates

VS





Benchmarking

Against other **methods**
(usually, using **metrics**)

Data Processing

Well, y'all know this already!

Problem to Solve

What are we implementing
(... and why?)

Benchmarking

Against other **biomarkers**
(... **more statistics**!) 🤓🤓

Hypothesis Testing

Lots of **statistics** 🤓

Data Processing

Well, y'all know this already!

Hypothesis (clinical)

What are we looking for
and why (**biomarkers**)

Liver fat from CT images predicts overall-survival in lung-cancer screening patients

Hypothesis (clinical)

What are we looking for and why (**biomarkers**)

Usually AI driven but not necessarily/only

Either develop from scratch or build on something already published

Data Processing

Well, y'all know this already!

Benchmarking

Against other biomarkers (... **more statistics**!)

Hypothesis Testing

statistics

Hypothesis (clinical)

What are we looking for and why (**biomarkers**)

Benchmarking

Against other biomarkers (...
more statistics!) 🤓🤓

Well, y

Show that the developed method works adjusting for **clinical covariates** (study and dataset -specific), subgroup analyses etc.

Hypothesis Testing

Lots of **statistics** 🤓

Hypothesis (clinical)

What are we looking for and why (**biomarkers**)

Benchmarking

Against other **biomarkers**
(... **more statistics**!) 🤓🤓

Show that the developed method works adjusting for **known biomarkers** (study and dataset -specific), subgroup analyses etc.

Hypothesis Testing

Data Processing

Well, y'all know this already!

Large Datasets

Varied Datasets

Real Datasets

Imaging Data + Clinical Data

Medical

- Hypothesis testing and generating
- Technical novelty is cool but we don't really care as long as it (really) works
- Really long times at first (basically building on top of the technical, so...), but easier to extend later
- Reviewers will look more at the scientific soundness of the paper
  - Hypotheses (clinical problem)
  - Generalisability
  - Statistics

Engineering

- Problem solving
- Technical novelty is usually very important (if not the only thing that actually matters)!
- Shorter times but harder to extend later (usually more novelty needed)
- Reviewers mostly care about improvements over SOTA
  - Not a lot of statistics (if any)
  - One dataset is usually ok
  - No use-cases are usually ok

Medical

Compare to gold standard (validation)

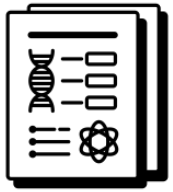
+

Downstream task(s):

- Liver segmentation for fat quantification (thresholding)
- Coronary Artery Calcium (CAC) segmentation (and quantification)
- Gamma analysis (RadOnc)
- Simpler biomarkers (e.g., the relative size of heart chambers from contrast CT has prognostic power [...])

Engineering

Segmentation



- Compare to gold standard (benchmark)
- Compare to (re-implemented) SOTA
- Ablation studies

Medical

Engineering

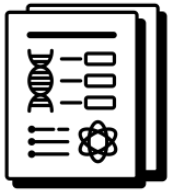
Compute AUROC

+

Downstream task(s):

- Statistical analysis
- Apply the model for eligibility in a trial (retrospective), see how things change
- Task-dependent clinical study
  - Investigate how the AI score correlates with other -omics data

Diagnosis/Classification



- Compute AUROC
- Ablation studies

Medical

Compute task-specific metrics
(e.g., SSIM, MAE, etc.) + reader test

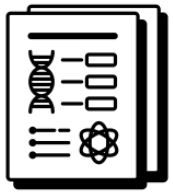
+

Downstream task(s):

- Re-planning
- Gamma analysis
- HU analysis
  - OOD data
  - Different Phantoms

Engineering

Image Translation (GANs)



- Compute task-specific metrics
  (e.g., SSIM, MAE, etc) + reader test
- Compare to (re-implemented) SOTA
- Ablation studies

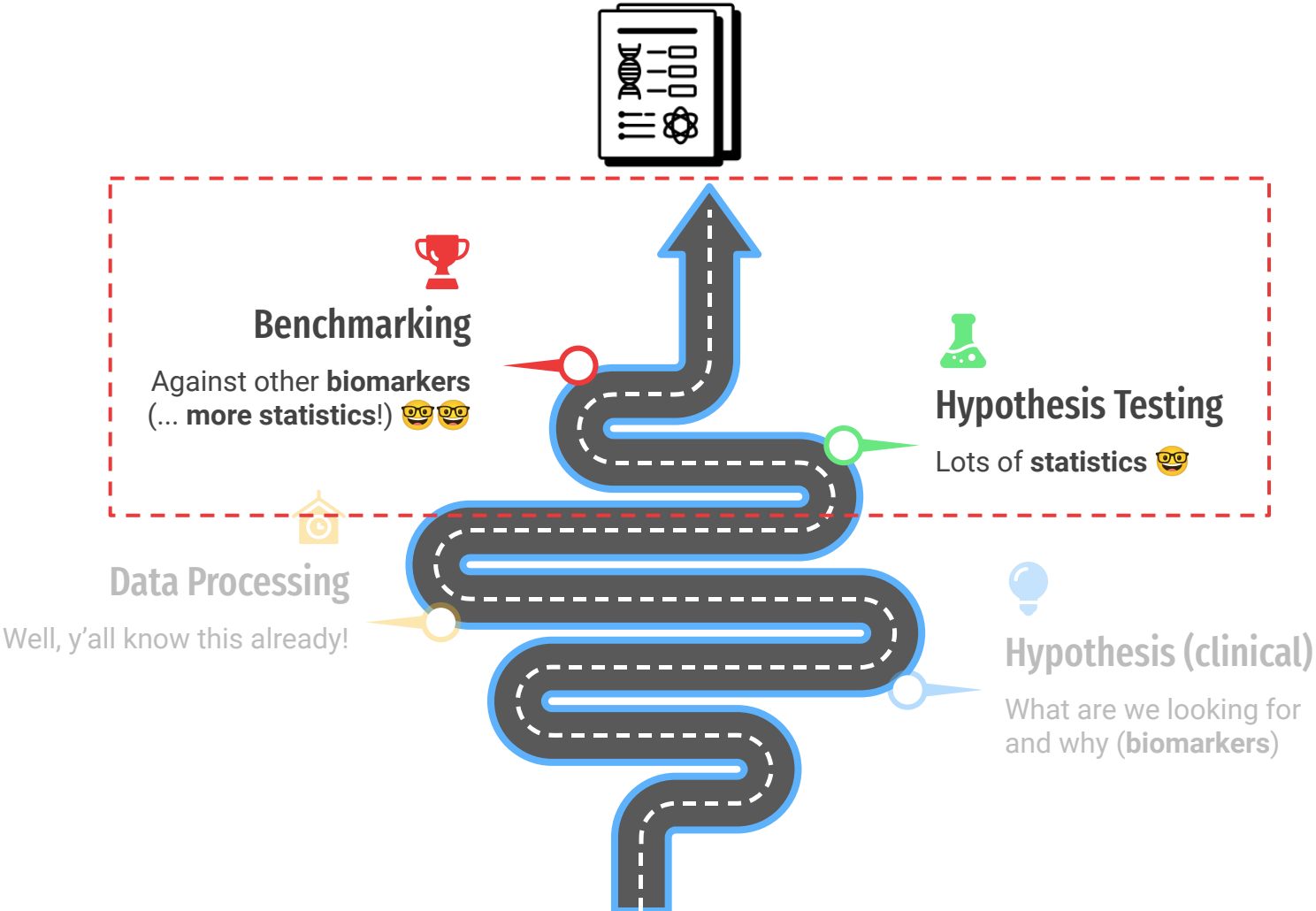

Benchmarking
Against other **biomarkers**
(... **more statistics**!) 🤓🤓
Hypothesis Testing
Lots of **statistics** 🤓
Data Processing
Well, y'all know this already!
Hypothesis (clinical)
What are we looking for
and why (**biomarkers**)

Table 2 Univariate and multivariable survival analyses of the predictive value of deep learning risk … ssed in the test cohorts.

Risk groups	Events[a]						P value	HR	95% CI	P value
NLST: n = 14,959; events: ASCVD death, n = 388								Adjusted for age, sex, diabetes, hypertension, past heart-disease, past stroke		
Very l								n/a	Reference	Reference
Low								1.57	0.96–2.57	0.069
Model								2.79	1.70–4.57	<0.001
High								3.87	2.45–6.11	<0.001
PROM										
Very l								Adjusted for Framingham Risk Score		
Low								n/a	Reference	Reference
Model									1.15–3.15	0.012
High									1.47–4.50	0.001
ROMI										
Risk g										

Martingale resi

Supplementary Table 3 Comparison …

Cohort							
FHS-CT2, n=663							
NLST, n=396							
PROMISE, n=4,021							
ROMICAT-II, n=441							
All test data, n=5521							

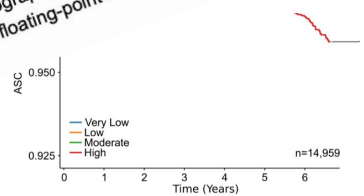
FHS-CT2[2]: Framingham Heart Study, paru… Screening Trial; PROMISE[4]: Prospective Multic… using Computer Assisted Tomography II; ICC: Intrac… which is the smallest positive floating-point number poss…

Supplementary Table 5 AUC comparison for event prediction based on automatically and manually calculated calcium scores.

		AUC	SD	SE	95%CI	P-value
NLST: n=396; Events: ASCVD death, n=11	AI	0.69	0.08	0.08	0.54-0.84	0.0156
	Manual	0.64	0.07	0.07	0.50-0.78	0.0552
PROMISE: n=4,021; Events: ACM, MI, UA, n=130	AI	0.66	0.02	0.02	0.62-0.71	3.63×10^{-11}
	Manual	0.68	0.02	0.02	0.63-0.72	9.34×10^{-13}
ROMICAT-II: n=441; Events: ACS, n=38	AI	0.83	0.03	0.03	0.76-0.90	1.96×10^{-14}
	Manual	0.86	0.03	0.03	0.81-0.92	1.94×10^{-16}

NLST[3]: National Lung Screening Trial; PROMISE[4]: Prospective Multicenter Imaging Study for Evaluation of Chest Pain; ROMICAT-II[5]: Rule Out Myocardial Infarction using Computer Assisted Tomography II; AUC: Area under the curve, SD: Standard Deviation, SE: Standard Error, CI: Confidence interval. All p-values were calculated using a two-sided Wilcoxon Test.

	Kappa
	0.726
	0.618



Number at risk

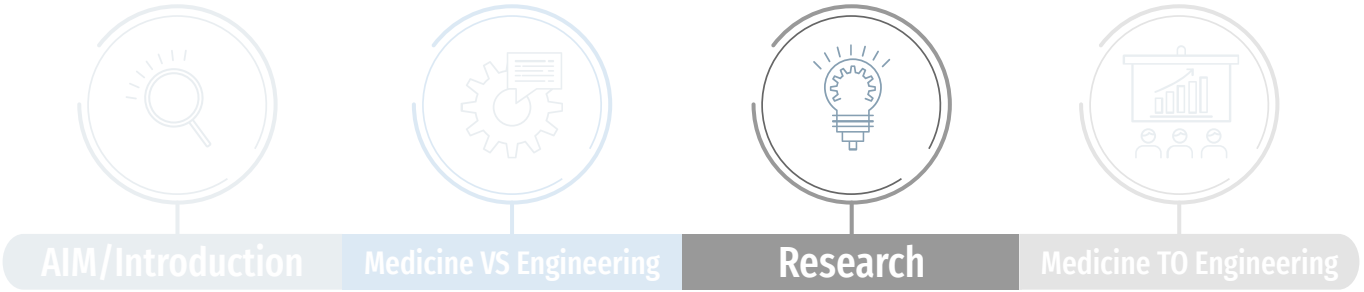
Very Low	3613	3596	3577	3532	3496	3451	2961
Low	4730	4706	4665	4614	4552	4492	3820
Moderate	2373	2352	2320	2294	2254	2216	1872
High	4243	4190	4129	4054	3955	3847	3231

Number at risk

Very Low	1615	1498	942	342
Low	1264	1150	713	234
Moderate	557	507	312	96
High	585	530	325	107

CAC group	Very Low	Low	Moderate	High
ACS=0 [n(%)]	256 (98.5)	105 (89.0)	27 (79.4)	15 (51.7)
ACS=1 [n(%)]	4 (1.5)	13 (11.0)	7 (20.6)	14 (48.3)
Total	260	118	34	29

Figure 4: Schoenfeld residuals plot for the multiv… cigarettes per day) Cox PH model.


AIM/Introduction
Medicine VS Engineering
Research
Medicine TO Engineering
What does the group focus on
What's different
Some of our research
What translates

DL in Medical Imaging

Data Mining

Mine the unknown.

New biomarkers discovery

Diagnostics

Quantify the known.

From qualitative to quantitative

Automation

Automate extraction.

Well known biomarkers extraction automation

DL in Medical Imaging

Data Mining

Mine the unknown.

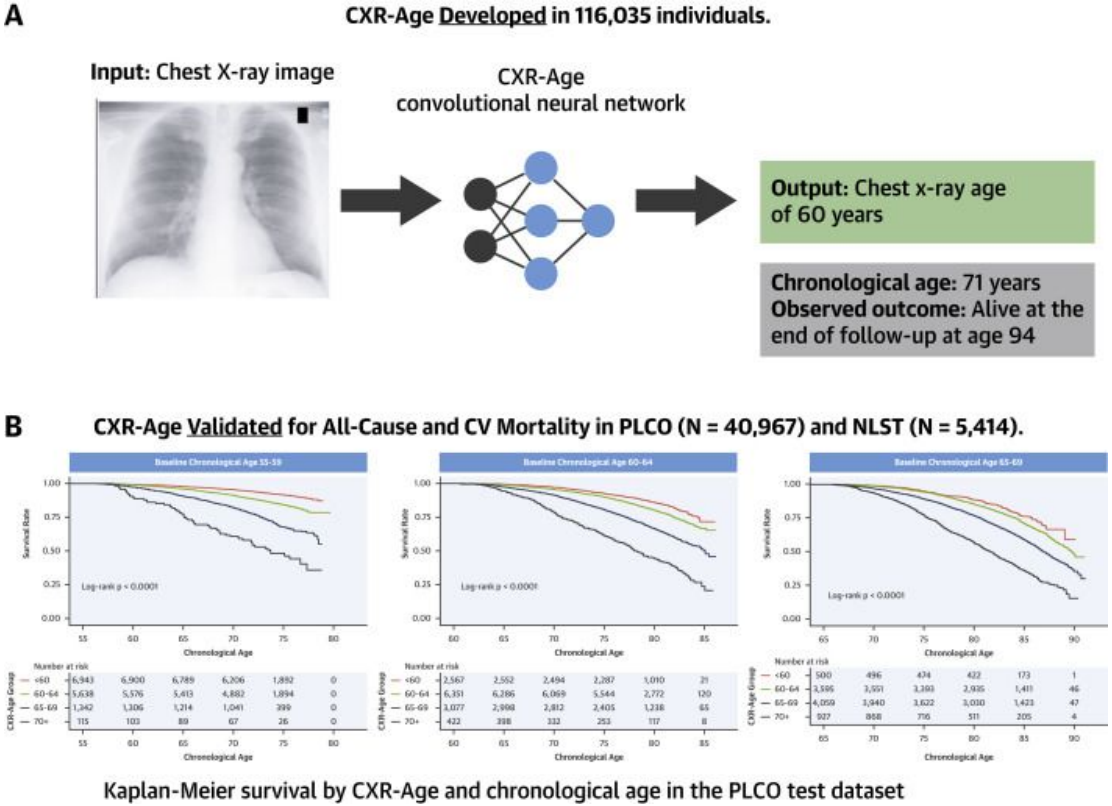
New biomarkers discovery

Try to extract information doctors don't really know how to mine (new biomarkers, always based on loose biological hypotheses anyway)

E.g., from Chest X-Rays:

- Deterioration of a patient with COVID (FB AI)
- Screening eligibility of heavy smokers

→ **hypothesis generating studies**



Vineet K. Raghu, Jacob Weiss, Udo Hoffmann, Hugo J. W. L. Aerts and Michael T. Lu.

Deep Learning to Estimate Biological Age From Chest Radiographs (JACC, 2021)

DL in Medical Imaging

Try to quantify qualitative information doctors use for clinical decision making (based on strong biological hypotheses)

E.g., from CT scans:

- Predict treatment response
- Predict survival of patients with lung lesions

From Chest X-Rays:

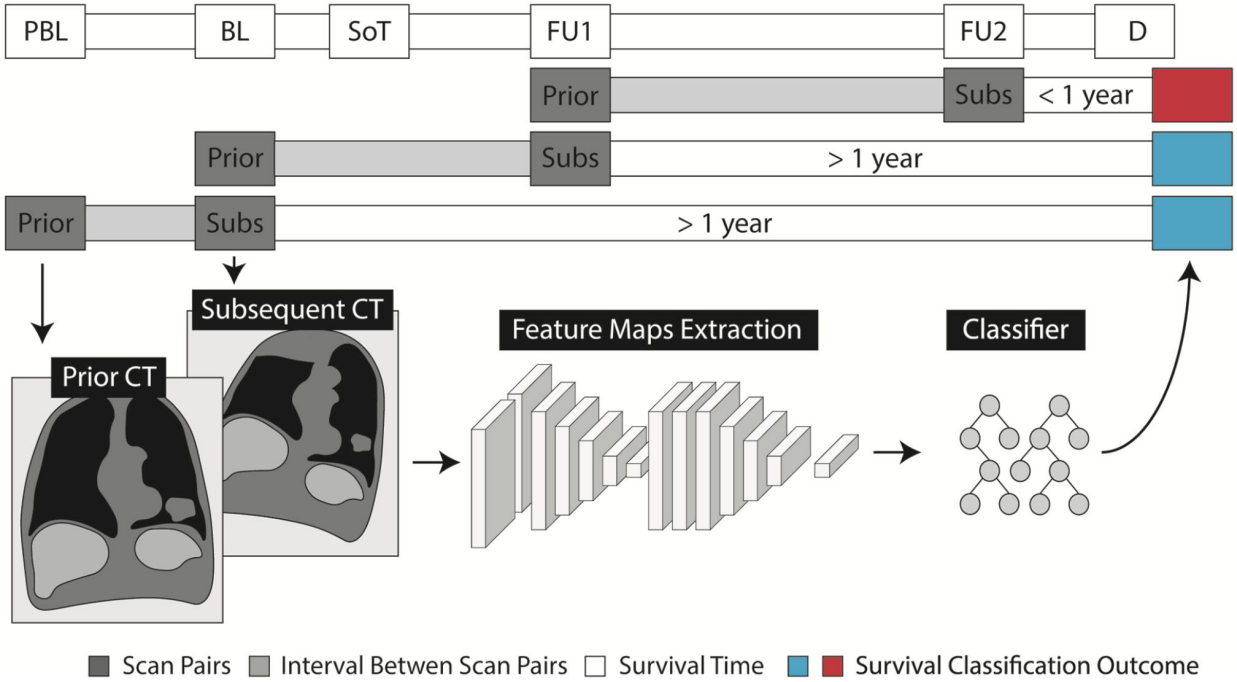
- COVID diagnose/severity

Diagnostics

Quantify the known.

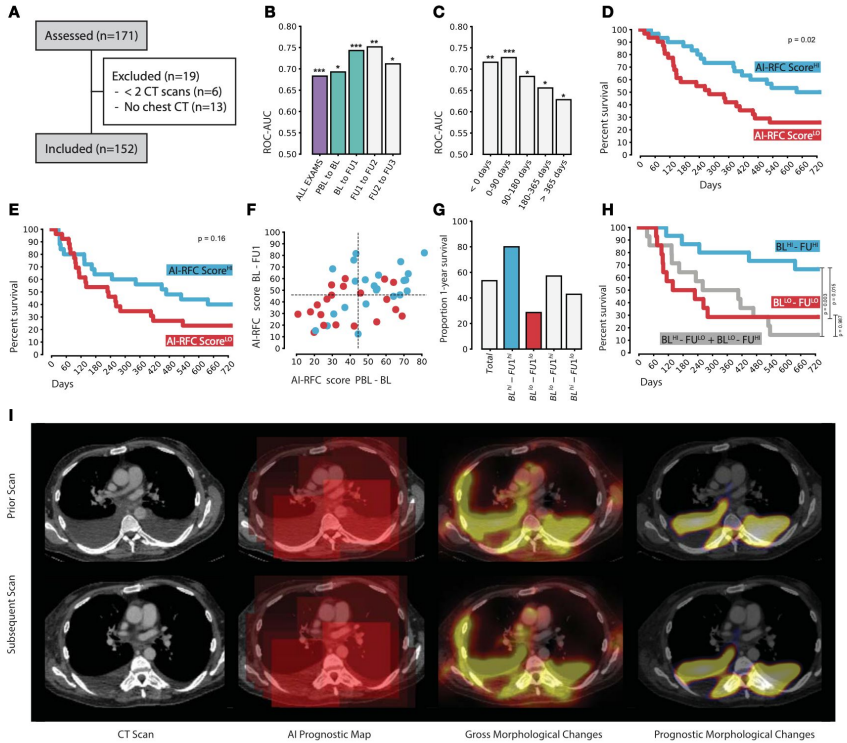
From qualitative to quantitative

→ **hypothesis testing studies**



Stefano Trebeschi, Zuhir Bodalal, Thierry N. Boellaard, Teresa M. Tareco Bucho, Silvia G. Drago, Ieva Kurilova, et Al.

Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy (Front. Oncol, 2021)



Stefano Trebeschi, Zuhir Bodalal, Thierry N. Boellaard, Teresa M. Tareco Bucho, Silvia G. Drago, Ieva Kurilova, et Al.

Prognostic Value of Deep Learning-Mediated Treatment Monitoring in Lung Cancer Patients Receiving Immunotherapy (Front. Oncol, 2021)

DL in Medical Imaging

Automate the extraction of information doctors know how to quantify (well known biomarkers) but often cannot due to time constraints

E.g., from CT scans:

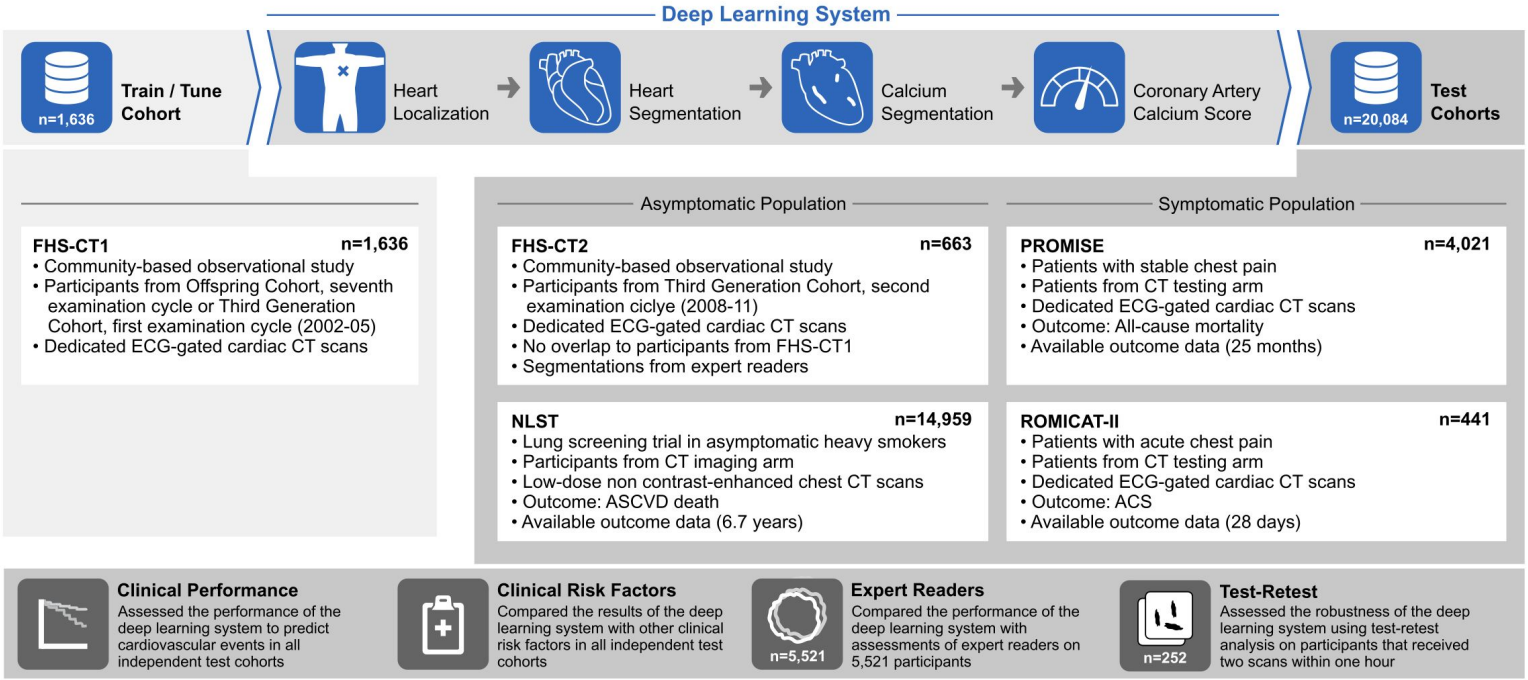
- Liver fat quantification
- Coronary Artery Calcium segmentation

→ **clinical integration, biomarker validation studies**

Automation

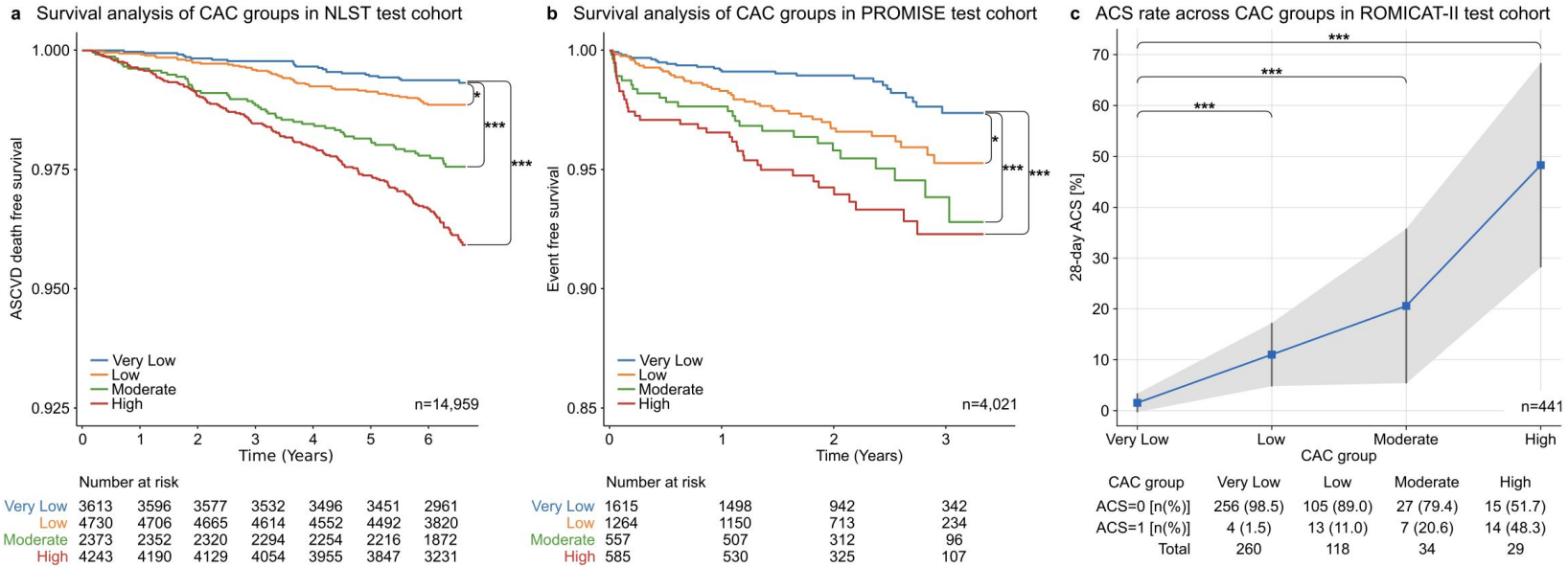
Automate extraction.

Well known biomarkers extraction automation



Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Ivanov Alexander, Jana Taron, Chintan Parmar, et Al.

Deep convolutional neural networks to predict cardiovascular risk from computed tomography (NatCom, 2021)



Roman Zeleznik, Borek Foldyna, Parastou Eslami, Jakob Weiss, Ivanov Alexander, Jana Taron, Chintan Parmar, et Al.

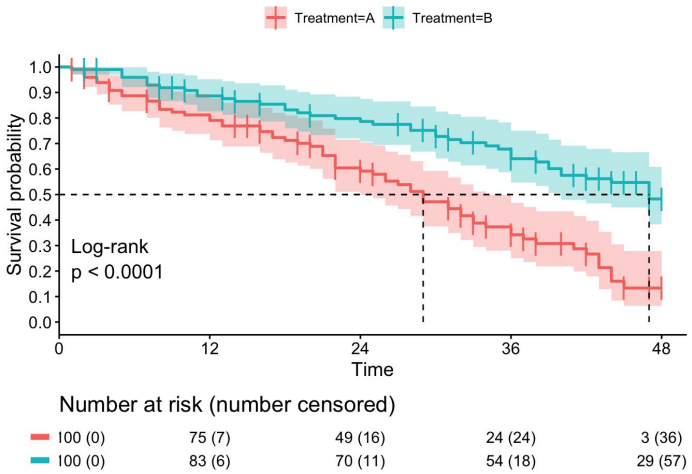
Deep convolutional neural networks to predict cardiovascular risk from computed tomography (NatCom, 2021)


AIM/Introduction
Medicine VS Engineering
Research
Medicine TO Engineering
What does the group focus on
What's different
Some of our research
What translates

bad
v
“There are three kinds of lies: lies, damned lies, and statistics”

Survival analysis techniques in SP/Network Engineering!

Kaplan-Meier



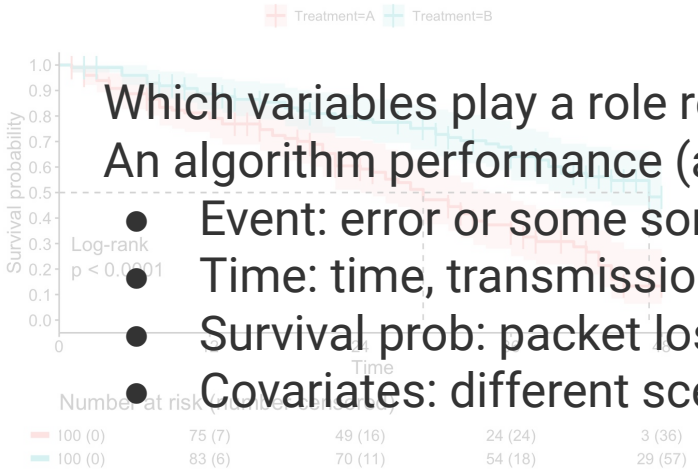
Cox Regression

Algorithm A VS Algorithm B

- Event: error of some sort
- Time: time, transmissions, distance
- Survival prob: packet loss

Survival analysis techniques in SP/Network Engineering!

Kaplan-Meier



Which variables play a role regarding
An algorithm performance (and how much)

- Event: error or some sort
- Time: time, transmissions, distance
- Survival prob: packet loss
- Covariates: different scenarios/setups

Cox Regression

	Recurrence only	
Variable	**Hazard ratio (95% CI)**	***P***
Age group, y		
65–69	1.00	
70–74	0.91 (0.87–0.97)	0.001
75–79	0.70 (0.66–0.74)	<0.001
AJCC stage		
I	1.00	
II	1.27 (1.21–1.33)	<0.001
III	2.04 (1.88–2.22)	<0.001
ER/PR status		
Positive[a]	1.00	
Negative	1.17 (1.09–1.25)	<0.001
Unknown	0.91 (0.85–0.96)	0.002
Histologic grade		
Well	1.00	
Moderate	1.22 (1.12–1.31)	<0.001
Poor	1.30 (1.20–1.41)	<0.001
Unknown	1.12 (1.03–1.21)	0.007