EUROPEAN HEMATOLOGY ASSOCIATION

Big data in cancer research: dangers and opportunities

Ton Coolen

Radboud University & Saddle Point Science Europe

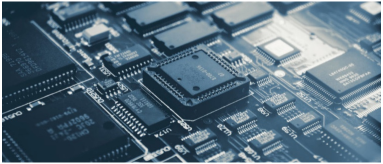
The Hague, February 9th 2024

What is ‘Big Data’?

- A: many samples, relatively few variables per sample

practical problems

(solved by larger disks,
faster computers,
parallelization of
existing algorithms)



- B: many variables per sample, relatively few samples

conceptual problems

– lack of intuition
– lack of appropriate methods

genomic data, images, ...

here conventional multi-variate methods
break down due to overfitting

Precision Cancer Medicine

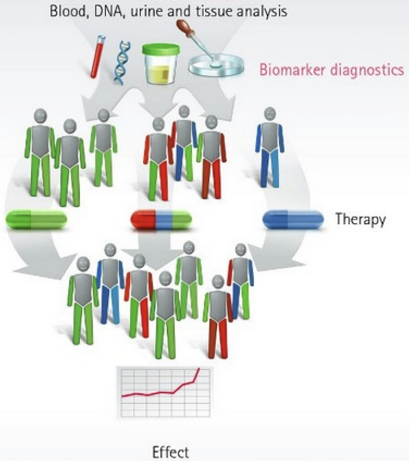
deep characterization of patients
in order to personalize therapy

- data with thousands or more measurements per patient
- but usually not with even *larger* numbers of patients

so: big data type B ...

(more measurements than samples,
overfitting danger)

we cannot yet use these data
fully and reliably without new methods ...



Precision Cancer Medicine

map *latent heterogeneity*
in diseases and their hosts

- identify drug responder subgroups,

  distinct in treatment associations?
  distinct in time courses?

impact of *ageing populations*

- interacting co-morbidities,

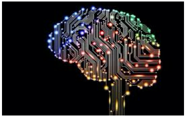
  decontaminate inferences for
  false aetiology/protectivity

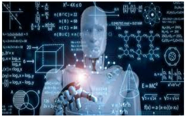
- longitudinal survival analysis

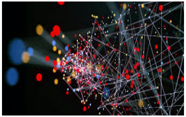
precision cancer medicine requires more complex statistical models
(making the sample size problem worse ...)

AI and Deep Learning

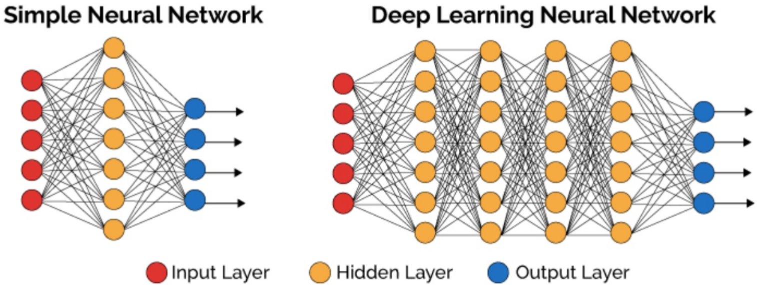
fancy names,
fancy pictures ...





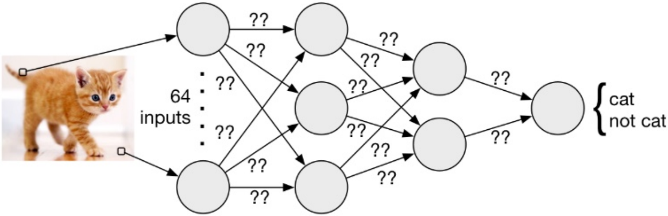


let's open the box:
1980s architectures, 1980s learning rules ...



Standard AI

- suitable problems



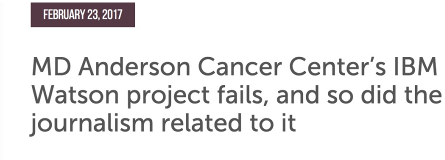
  – many data of the type (question + answer)
  – no need for explanations

  e.g. speech recognition, digital pathology

- limitations of AI in medicine

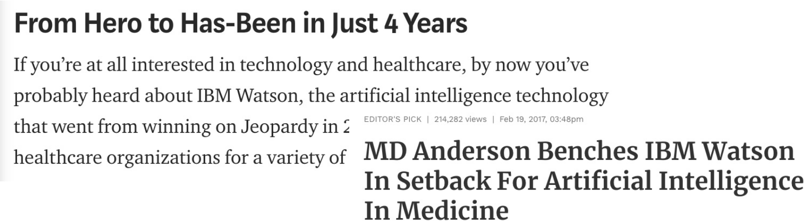
  – ‘black box’ decisions without reliable error bars
  – cannot handle complexities such as confounders, disease interactions, latent heterogeneity
  – no counterfactual reasoning

Dangers of AI hyping ...



FEBRUARY 23, 2017

MD Anderson Cancer Center's IBM Watson project fails, and so did the journalism related to it

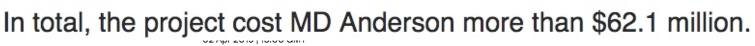


From Hero to Has-Been in Just 4 Years

If you're at all interested in technology and healthcare, by now you've probably heard about IBM Watson, the artificial intelligence technology that went from winning on Jeopardy in 2 healthcare organizations for a variety of

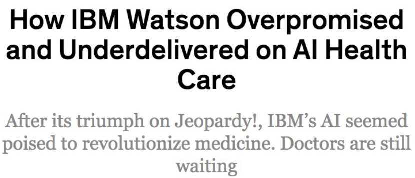
EDITOR'S PICK | 214,282 views | Feb 19, 2017, 03:48pm

MD Anderson Benches IBM Watson In Setback For Artificial Intelligence In Medicine



In total, the project cost MD Anderson more than $62.1 million.

02 Apr 2019 | 13:00 GMT



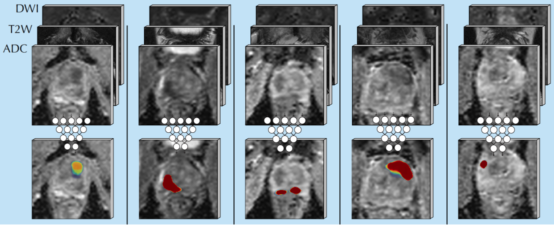
How IBM Watson Overpromised and Underdelivered on AI Health Care

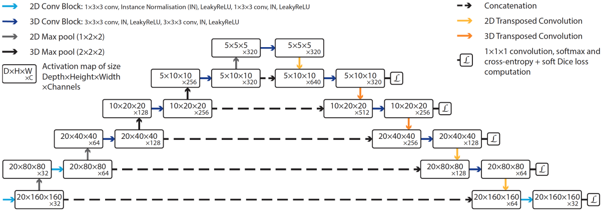
After its triumph on Jeopardy!, IBM's AI seemed poised to revolutionize medicine. Doctors are still waiting

Main success stories of AI in medicine

- segmentation and feature detection in clinical images

  – as accurate as humans
  – but massively faster and cheaper





2D Conv Block: 1×3×3 conv, Instance Normalisation (IN), LeakyReLU, 1×3×3 conv, IN, LeakyReLU

3D Conv Block: 3×3×3 conv, IN, LeakyReLU, 3×3×3 conv, IN, LeakyReLU

2D Max pool (1×2×2)

3D Max pool (2×2×2)

DxHxW ×C Activation map of size Depth×Height×Width ×Channels

Concatenation

2D Transposed Convolution

3D Transposed Convolution

ℒ 1×1×1 convolution, softmax and cross-entropy + soft Dice loss computation

Corollary

- modern cancer research needs new quantitative tools
  - sample size problems
  - complexities of heterogeneous and elderly populations
  - interpretable
- AI is excellent in digital pathology
  - (so far) unable to deal with above challenges
  - but can inspire new statistical algorithms ...

Statistical innovation for cancer research

- unfortunately very slow ...
  - journals discourage non-standard methods (‘our readership ...’)
  - who writes the industry-standard user-friendly code?
    (no programmers in stats departments → spin-outs)
  - epidemiologists too busy with routine tasks
  - statisticians see limited benefit in reaching out

Proposals for analytical innovation in cancer research

for which validated methodology already exists!

1. Include more covariates / do more with fewer samples

   – overfitting correction methods
   – federated Bayesian inference

2. Longitudinally updated personalized survival prediction

   – being alive later changes survival curves, even without involving data

3. Inference of personalized optimal treatment dose

   – via interaction terms in existing survival analysis models

4. Correct predictions for interacting comorbidities

   – decontaminated survival curves
   – decontaminated associations and hazard ratios

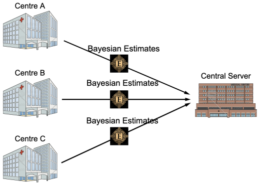
5. Identification of responders in phase 2 or 3 cancer trials

   – more options for patients via rescue of failed trials
   – prevention of pointless side effects
   – better use of cancer research funds

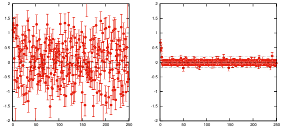
Remainder of this talk:

examples of new quantitative tools for cancer research

- Bayesian Federated Inference (BFI)



- Overfitting correction methods and pipelines



- Responder subgroup identification in cancer trials

Bayesian Federated Inference

harness the power of large datasets *without creating large data sets*

The problem

multivariate analysis requires
large data sets to avoid *overfitting*

rare diseases: always small data sets ...

Possible solutions

1. more effective mechanisms and
   incentives for data sharing

2. technology for integration of
   individual analysis outcomes

   reconstruct from local analyses on data subsets what would have been found if these had been combined into a single larger data set

Centre A

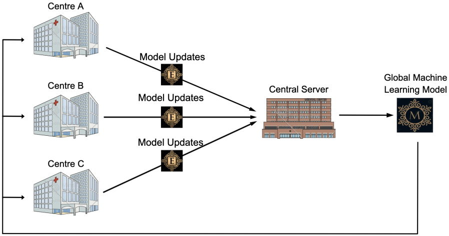
Centre B

Centre C

2017: Federated Machine Learning

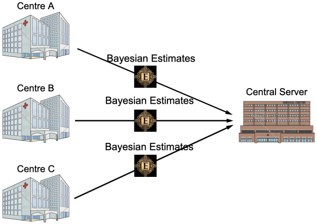
disadvantages

- many iterations needed
- complex infrastructure
- labour intensive
- data security difficult to control
- black box algorithms
- predictions without error bars



2020: Bayesian Federated Inference

- only one (more complex) analysis needed
- no convergence issues
- no data security issues
- fully interpretable statistical models
- predictions with error bars



Pilot tests of BFI on real data

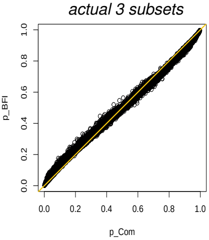
trauma patients from different hospitals,
4 covariates, outcome: death (yes/no)

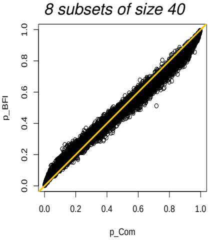
data subsets	size	mortality	age	gender	ISS	GCS
	n_ℓ	%	median	% females	median	median
peripheral hospitals without NSU	49	43	42	22	41	11
peripheral hospitals with NSU	106	40	34	24	33	14
academic hospitals	216	22	35	30	29	11
combined data	371	30	36	27	30	12

(NSU: neuro-surgical unit)

death probabilities:

combined set (p_Com) versus
BFI-reconstructed (p_BFI)


actual 3 subsets


8 subsets of size 40

Ongoing BFI research

how to handle protocol differences between centres

compare two chemotherapies, A and B,
using data from two medical centres

	CHEMO A	CHEMO B
medical centre 1	40% (40/100)	30% (150/500)
medical centre 2	18% (36/200)	15% (12/80)

both centres agree:
A is better

now combine our data!

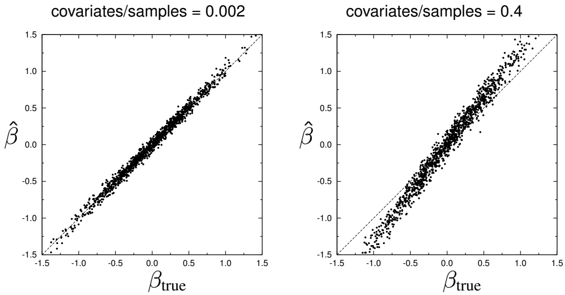
	CHEMO A	CHEMO B
medical centre 1	40% (40/100)	30% (150/500)
medical centre 2	18% (36/200)	15% (12/80)
response rate	25% (76/300)	28% (162/580)

are we still sure?
(Simpson’s paradox)

Overfitting Correction Methods and Pipelines

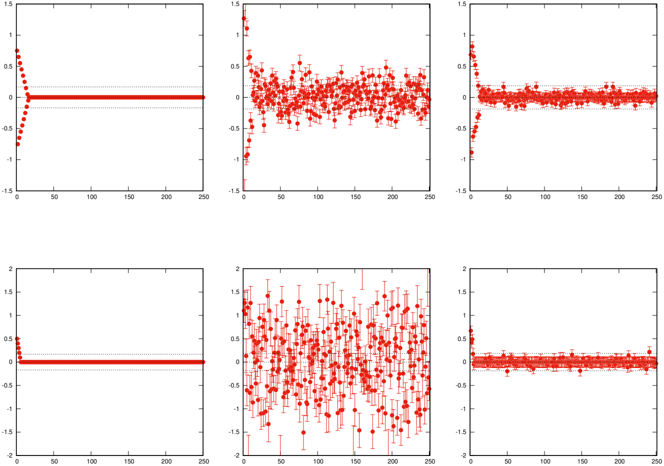
based on *mathematical understanding* of overfitting

Cox-inferred versus true association parameters
(simulated survival data)



- effect on regression parameters: inflation + noise
- both can be predicted mathematically,
  → correction formulae → fewer samples needed

example: 400 samples,
250 covariates (of which only a few informative)



true associations

Cox regression

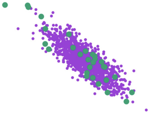
corrected

uncorrelated covariates

correlated covariates

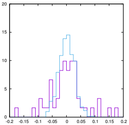
Automated pipeline: SaddlePoint Signature

preprocessing



normalization, imputation
informative missingness,
multiplexing

covariate pre-selection

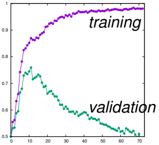


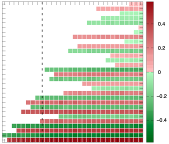
univariate regression,
correlation with outcome,
relative to randomized

main regression loop

backward iteration, 10^4 randomizations each
L_2 priors with adaptive hyperparameters
replica theory overfitting correction
probabilistic removal criterion

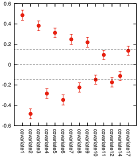
covariate selection





optimal covariate set

robust signatures

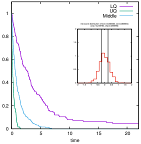


```
multivariate risk score formula:

S=(0.164947)*WHO
+ (−0.231909)*TSTAT:Resected
+ (0.001625)*SUMLES
+ (−0.062253)*nEREG
+ (0.412737)*RAS:Mutation
+ (0.627957)*BRAF:Mutation
+ (0.028880)*NEUT
+ (0.000688)*ALKP
+ (0.226668)*SPAIN0
− (1.028487)
```

prognostic score (incl covariate interactions)
treatment response score
probabilistic outcome predictions

visualize stratification



survival curves of risk
or response score quartiles,
ROC curves, score distributions ...

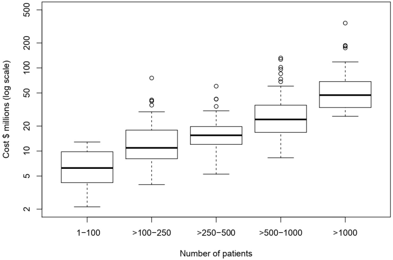
Responder subgroup identification

who actually benefits from a cancer drug?
prevent and rescue failed trials

The problem

poor drug targeting

– more failed clinical trials
– fewer treatment options for patients
– pointless side effects
– enormous waste of resources



- *phase 2 trials:*

  costs ∼15M$
  success rate 50% (cancer 33% ...)

- *phase 3 trials:*

  costs ∼30M$
  success rate 60% (cancer 36% ...)

Responder subgroups in failed cancer trials

weak drug benefit, no license ...
(in absence of response biomarker)

Two possibilities

1. *reproducible individual response*

   there are measurable differences between individuals that explain response variation, we just don’t know what they are ...

   cohort is stratifiable, drug can be rescued

2. *non-reproducible individual response*

   there are no measurable differences between individuals to explain response variation

   cohort is not stratifiable, drug cannot be rescued

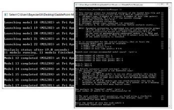
Bayesian latent class survival analysis

- reports characteristics of latent strata
- fully interpretable
- retrospective stratification: tool for finding subgroup markers
- prospective stratification *if covariates informative*

Automated pipeline: SaddlePoint Mosaics

Multi-risk latent class analysis / Regression management

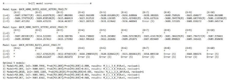
Automated regression management



```
Launching model 18 (M1L3K2) at Fri Ap
Launching model 19 (M2L3K2) at Fri Ap
Launching model 20 (M3L3K2) at Fri Ap
Launching model 21 (M1L3K3) at Fri Ap
Launching model 22 (M2L3K3) at Fri Ap
Analysis status after 19.0 seconds:
11 models running, 11 models finished.
-----------------
Model 12 completed (M3L2K3) at Fri Ap
Model 13 completed (M1L2K4) at Fri Ap
Model 14 completed (M2L2K4) at Fri Ap
Model 15 completed (M3L2K4) at Fri Ap
Model 16 completed (M1L3K1) at Fri Ap
Model 17 completed (M2L3K1) at Fri Ap
```

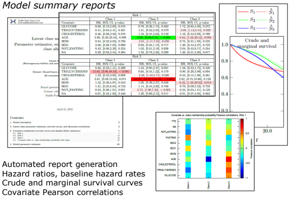
Automated regression and model ‘likelihood’ score determination for all candidate models
Baseline hazard rate(s), covariate association, and frailty estimation for each model
Bayesian priors on all parameters

Bayesian model selection



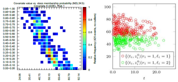
Automated Bayesian model likelihood determination and model ranking
Aikake and Bayesian Information Criterion scores also available

Model summary reports



Automated report generation
Hazard ratios, baseline hazard rates
Crude and marginal survival curves
Covariate Pearson correlations

Retrospective class allocation

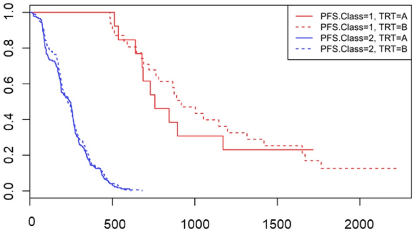


Allocate patients to most probable latent class
Data-driven cohort stratification
Covariate and class membership correlations
Aids search for new informative biomarkers

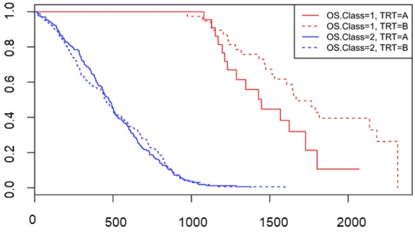
The COIN trial (colorectal cancer)

$n = 398,\ 1630$

PFS



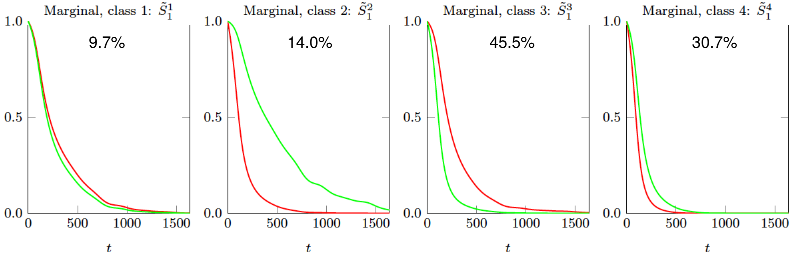
OS



The TOPICAL trial (lung cancer)

$n = 580$

	Risk 1			
	Class 1	Class 2	Class 3	Class 4
Covariate	HR, 95% CI, *p*-value	HR, 95% CI, *p*-value	HR, 95% CI, *p*-value	HR, 95% CI, *p*-value
AGE	0.77, [0.34,1.72], 0.521	2.93, [1.49,5.73], 0.002	0.59, [0.17,1.98], 0.390	0.92, [0.45,1.87], 0.819
Male	0.79, [0.36,1.74], 0.560	1.78, [0.92,3.42], 0.086	0.88, [0.33,2.35], 0.806	3.12, [1.44,6.74], 0.004
ECOG 2-3	0.40, [0.13,1.19], 0.099	1.49, [0.86,2.57], 0.156	1.54, [0.81,2.94], 0.186	1.75, [0.72,4.28], 0.216
Stage IV	1.34, [0.75,2.39], 0.326	1.46, [0.80,2.67], 0.219	1.96, [0.85,4.55], 0.116	1.20, [0.67,2.15], 0.539
Adenocarcinoma	7.20, [2.61,19.85], < 0.001	0.44, [0.24,0.82], 0.009	1.46, [0.65,3.31], 0.361	0.68, [0.33,1.39], 0.291
Ex-smoker	2.04, [0.56,7.48], 0.281	0.19, [0.06,0.63], 0.006	8.39, [2.12,33.18], 0.002	0.69, [0.27,1.77], 0.438
Smoker	5.04, [1.50,16.98], 0.009	0.30, [0.09,1.05], 0.060	4.99, [1.31,18.96], 0.018	1.19, [0.47,3.00], 0.717
CCI 4+	1.41, [0.65,3.06], 0.386	1.47, [0.80,2.67], 0.211	0.87, [0.25,2.96], 0.818	1.21, [0.55,2.63], 0.636
Good	0.32, [0.15,0.65], 0.002	0.23, [0.12,0.46], < 0.001	0.43, [0.21,0.87], 0.019	1.35, [0.70,2.61], 0.366
Tarceva	1.48, [0.79,2.76], 0.223	0.11, [0.05,0.22], < 0.001	3.95, [0.94,16.66], 0.061	0.45, [0.20,1.00], 0.050



survival curves: green=erlotinib, red=placebo

Thanks to

- collaborators NL:

  Emanuele Massa, Marianne Jonker,
  Hassan Pazira, Theodore Nikoletopoulos

- collaborators UK:

  Mark Rowley, Mieke van Hemelrijck, Alexander Mozeika,
  Fabrizio Antenucci, Paul Barber

- funding:

MRC Medical Research Council

CANCER RESEARCH UK

HANARTH FONDS

NWO

Papers, presentations, software

a.coolen@science.ru.nl
ton.coolen@saddlepointscience.com
https://toncoolen.wixsite.com/accc