FAIR data-sharing

Federated learning in Healthcare

Erik Roelofs, PhD
Medical Physicist
Manager Medical Informatics
Head of Software Development Team

Department of Radiation Oncology (MAASTRO)
GROW - Maastricht University Medical Centre +
Maastricht, The Netherlands

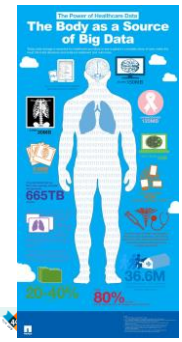
Disclosures

- 50% MAASTRO clinic; 50% Maastricht University *(→ptTheragnostic)*
- Research collaborations/funding
  - Varian (VATE, chinaCAT, euroCAT), Siemens (euroCAT), Sohard (SeDI, CloudAtlas), Mirada Medical (CloudAtlas), Philips (EURECA, TraIT, SWIFT-RT), Xerox (EURECA), De Praktijkindex (DLRA), *ptTheragnostic/DNAMito* (CTO: *PRODECIS*)
- Public research funding
  - Radiomics (USA-NIH/U01CA143062), euroCAT(EU-Interreg), duCAT (NL-STW), EURECA (EU-FP7), SeDI & CloudAtlas (EU-EUREKA), TraIT (NL-CTMM), DLRA (NL-NVRO)

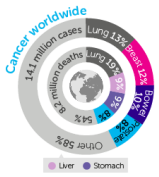
Data driven healthcare

Big Data in healthcare

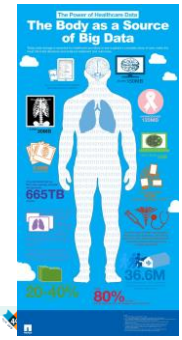
- 2/3 petabytes in average hospital
  (1500 year of music on a mobile phone)
- Growth 20-40% per year
- 80% unstructured
  - mostly imaging
  - diverse origin (CT, MR, US, 2D, 3D, 4D)



Big Data?



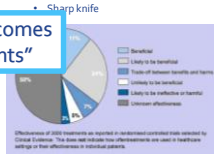
Oncology
2005-2015
140M patients
0.1-10GB per patient
14-1400PB
80% unstructured
100k hospitals



The problem of Big Data – The doctor is drowning

"We cannot predict outcomes of individual treatments"

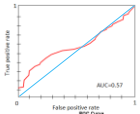
- Explosion of data
- Explosion of decisions
- Explosion of 'evidence'*
  - 3 % in trials, bias
  - Sharp knife



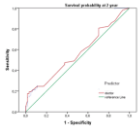
*2010: 1574 & 1354 articles on lung cancer & radiotherapy = 7.5 per day
Half-life of knowledge estimated at 7 years (in young students)

J Clin Oncol 2010;28:4268, JMI 2012 Friedman, Rigby, BMJ Clinical Evidence

Prediction by MDs?



NSCLC
2 year survival
30 patients
8 MDs
Retrospective
AUC: 0.57



NSCLC
2 year survival
158 patients
5 MDs
Prospective
AUC: 0.56

Oberije et al.

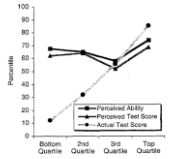


Figure 2. Perceived logical reasoning ability and test performance as a function of actual test performance (Study 2).

Unskilled and unaware of it: *How difficulties in recognizing one's own incompetence leads to inflated self-assessments.* J Pers Soc Psych

Kruger et al. 1999

Evidence-based medicine in Rad.Onc.?

- In radiotherapy we treat patients using complicated (rapidly-changing) technology
- Getting RCT evidence on technology is hard (*often late: e.g. protons*)
- This will become worse as our knowledge of cancer and technology increases

Future radiotherapy practice will be based on evidence from retrospective interrogation of linked clinical data sources rather than prospective randomized controlled clinical trials

Andre L. A. J. Dekker, Ph.D.
Knowledge Engineering Department, MAASTRO Clinic, Maastricht 6229 ET, The Netherlands
(Tel: 31 (0)88-4455666; E-mail: andre.dekker@maastro.nl)

Sarah L. Gulliford, Ph.D.
Department of Physics, Royal Marsden NHS Foundation Trust and The Institute of Cancer Research, Sutton, Surrey SM2 5PT, United Kingdom
(Tel: 020 7153320; E-mail: sarah.gulliford@icr.ac.uk)

Martin A. Ebert, Ph.D. and Colin G. Orton, Ph.D., Co-Moderators

(Received 13 October 2013; accepted for publication 14 October 2013; published 4 February 2014)

[http://dx.doi.org/10.1118/1.4832139]

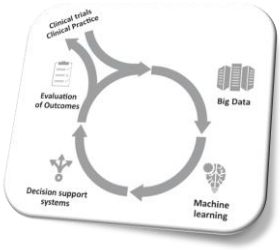
Rapid-learning healthcare

The potential of Big Data - Rapid Learning Health Care

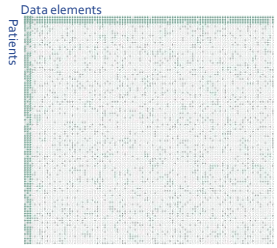
In [..] rapid-learning, [..] data routinely generated through **patient care and clinical research** feed into an ever-growing [..] set of **coordinated databases**.
J Clin Oncol 2010;28:4268

[..] rapid learning [..] where we can **learn from each patient** to guide practice, is [..] crucial to guide rational health policy and to contain costs [..].
Lancet Oncol 2011;12:933

Examples:
DLRA, CAT (www.eurocat.info)
ASCO's CancerLinQ



Data landscape



- Clinical research
  - 3% of patients
  - 100% of features
  - 5% missing
  - 285 data points
- Clinical registries
  - 100% of patients
  - 3% of features
  - 20% missing
  - 240 data points
- Clinical routine
  - 100% of patients
  - 100% of features
  - 80% missing
  - 2000 data points

Data Landscape – completing data

Data Landscape – who has what?

Data elements

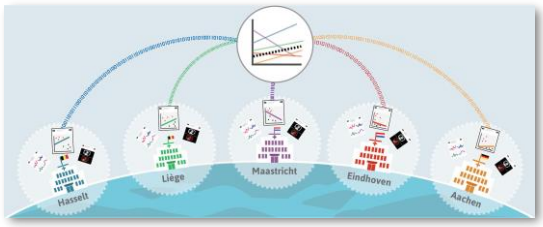
Patients

A different approach

- If sharing is the problem: don't share the data!
- If you can't bring the data to the learning application,

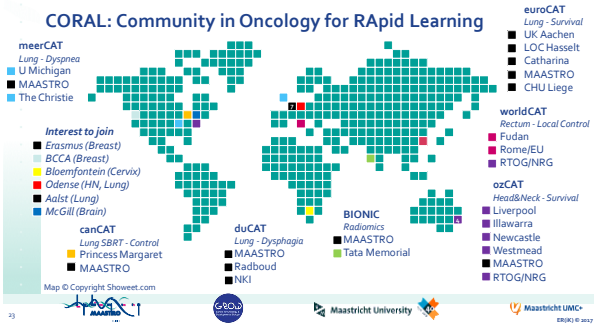
  you have to bring the learning application to the data
- Consequences
  - The learning application has to be distributed
    - A federated learning network.
  - The data has to be understandable by an application (i.e. computer, not a human)
    - Semantic Interoperable → FAIR*ify* medical databases

EuroCAT : distributed learning network



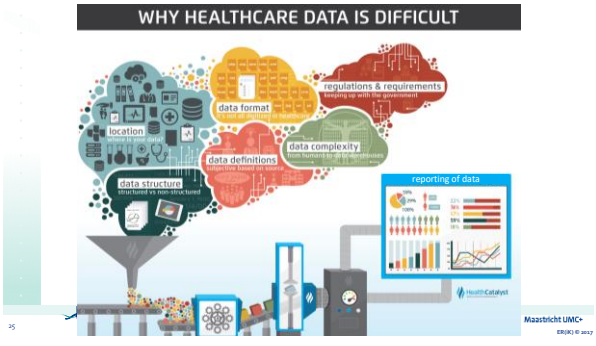
Deist, T.M., et al., doi:10.1016/j.ctro.2016.12.004

https://youtu.be/ZDJFOxpwqEA

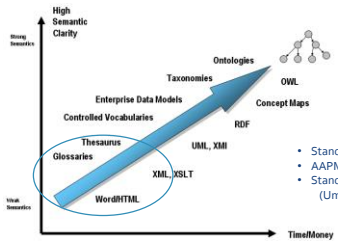


Data stewardship and management

Working towards FAIR data stations



Towards semantic web technology



- Standard naming conventions
- AAPM TG 263
- Standard data collection protocols (Umbrella's) : cancerdata.org/protocols

http://www.mkbergman.com/374/an-intrepid-guide´-to-ontologies/


SCIENTIFIC DATA nature.com/articles/sdata201618

The FAIR Guiding Principles for scientific data management and stewardship

Mark D. Wilkinson, Michel Dumontier [...] Barend Mons

Affiliations | Contributions | Corresponding author

Scientific Data 3, Article number: 160018 (2016) | doi:10.1038/sdata.2016.18

Received 10 December 2015 | Accepted 12 February 2016 | Published online 15 March 2016

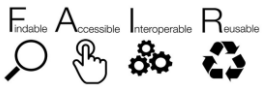
About Scientific Data

Scientific Data is an open-access, peer-reviewed journal for descriptions of scientifically valuable datasets. Our primary article-type, the Data Descriptor, is designed to make your data more discoverable, interpretable and reusable.

Principles to enhance the value of *all* digital resources and their metadata.

data, images, software, web services, repositories

FAIR data



- **F**indable: the data should be *uniquely* and *persistently identifiable* (PID) and other researchers should be able to *find* your data.
- **A**ccessible: the *conditions* under which the data can be used should be *clear* to machines and humans.
- **I**nteroperable: data should be *machine-readable* and use *terminologies*, *vocabularies*, or *ontologies* that are *commonly used* in the field;
- **R**eusable: compliant with the above and sufficiently well described with *metadata* and *provenance information* so that the data sources can be *linked* or *integrated* with other data sources and enable proper *citation*.

Rapid Adoption of Principles

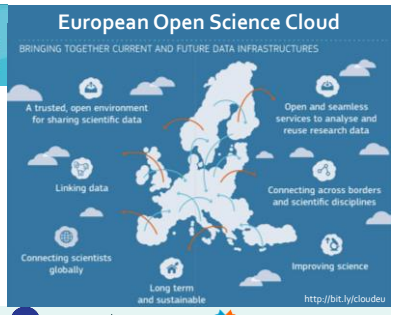
Developed and endorsed by researchers, publishers, funding agencies, industry partners.

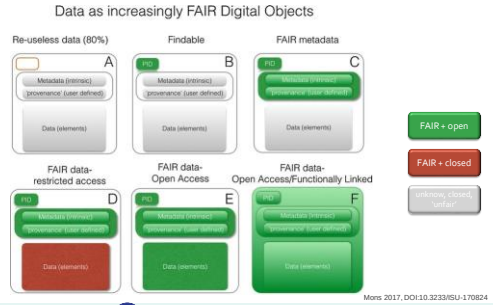
As of May 2017, 100+ citations since 2016 publication

Included in G20 communique, EOSC, H2020, NIH, and more…

European Cloud Initiative

Building a competitive data and knowledge economy in Europe

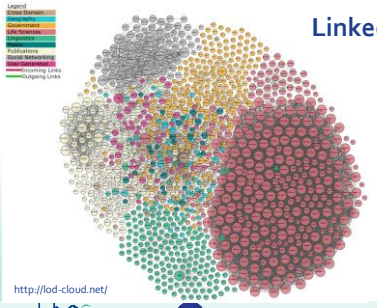




@micheldumontier::#DANSLOD:2017-05-01

Evaluation 40 repositories

Linked Open Data



LOD cloud diagram

overview of datasets published in Linked Data format, with rich metadata

http://lod-cloud.net/

SAM questions

- **Why are doctors bad at predicting outcomes?**
  - A. They get confused by too much data
  - B. They have too many options to choose from
  - C. There is not enough good evidence to make a choice
  - D. All of the above
- **What is the best way to validate a model?**
  - A. Bootstrap your training set to create a validation set
  - B. Send the model to another hospital and have them validate it
  - C. Get data from another hospital to validate
  - D. Split your own data before modelling into a training and validation set
- **What is generally the best way to improve models in machine learning?**
  - A. Remove outlying patients from the training dataset
  - B. Extend the training dataset with more patients
  - C. Extend the training dataset with more features
  - D. Use a better machine learning algorithm

"In God we trust. All others must bring data".

W. Edwards Deming

Maastricht University

Maastricht UMC+

SAM questions

- **Why are doctors bad at predicting outcomes?**
  - A. They get confused by too much data
  - B. They have too many options to choose from
  - C. There is not enough good evidence to make a choice
  - D. All of the above
- **What is the best way to validate a model?**
  - A. Bootstrap your training set to create a validation set
  - B. Send the model to another hospital and have them validate it
  - C. Get data from another hospital to validate
  - D. Split your own data before modelling into a training and validation set
- **What is generally the best way to improve models in machine learning?**
  - A. Remove outlying patients from the training dataset
  - B. Extend the training dataset with more patients
  - C. Extend the training dataset with more features
  - D. Use a better machine learning algorithm

"In God we trust. All others must bring data".

W. Edwards Deming

Maastricht University

Maastricht UMC+

Acknowledgements

- Varian, Palo Alto, CA, USA
- Siemens, Malvern, PA, USA
- RTOG, Philadelphia, PA, USA
- MAASTRO, Maastricht, Netherlands
- Policlinico Gemelli, Roma, Italy
- UH Ghent, Belgium
- Catherina Zkh Eindhoven, Netherlands
- UZ Leuven, Belgium
- Radboud, Nijmegen, Netherlands
- University of Sydney, Australia
- UMCG, Groningen, Netherlands
- Liverpool and Macarthur CC, Australia
- CHU Liege, Belgium
- Uniklinikum Aachen, Germany
- LOC Genk/Hasselt, Belgium
- Princess Margaret Hospital, Canada
- The Christie, Manchester, UK
- UH Leuven, Belgium
- State Hospital, Rovigo, Italy
- Illawarra Shoalhaven CC, Australia
- Fudan Cancer Center, Shanghai, China

Maastricht University

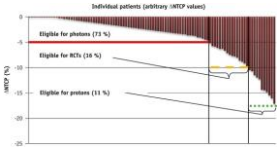
Maastricht UMC+

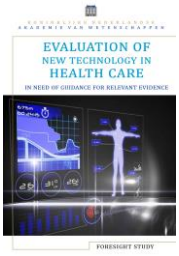
Thank you for your attention

FAIR adoption by European Commission

Model based approach

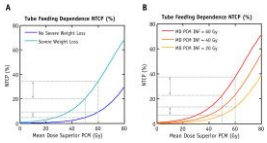
- Proton therapy introduction in the Netherlands
- Expensive and only 1800 slots
- ALARA -> protons for reduced toxicity
- RCT -> protons for better survival/control
- Evidence-based (e.g. paediatric) and model-based indications (HN, GBM, Lung, Breast, Prostate)





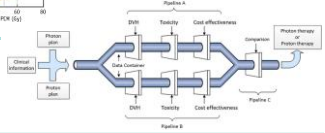
Widder et al. http://dx.doi.org/10.1016/j.ijrobp.2015.10.004

Model based approach



Widder et al. http://dx.doi.org/10.1016/j.ijrobp.2015.10.004

Cheng et al. http://dx.doi.org/10.1016/j.radonc.2015.12.019

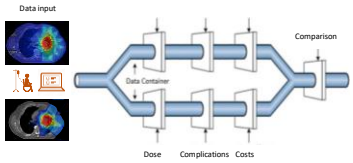


There is an app for that

- Endometrium cancer
- Head & neck cancer
- Brain metastases
- Lung cancer
- Prostate cancer
- Rectum cancer

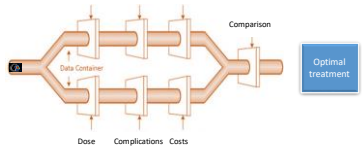
Model-based proton therapy

Proton Decision Support (PRODECIS)



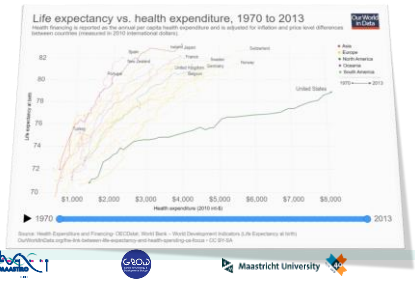
Model-based proton therapy

Proton Decision Support (PRODECIS)



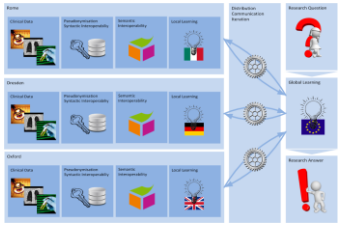
Personal Health Train

The example of … healthcare efficiency

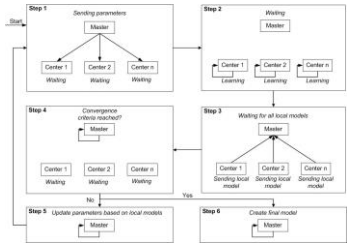


euroCAT basic setup

- Keep data locally
- Standardize it according to an ontology
- Make and send around learning (query & calculate) "bots"
- Learn prediction models for cancer
- Share the prediction model – not the data!



Distributed learning - math



RDF Graph example