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, TruT, SWIFT-RT), Xerox (EURECA), De Praktijkindex (DLRA), ptTheragnostic/GNMMS (CTO: PRODECIS)

• Public research funding
  – Radiomics (USA-NIH/U01CA143062), euroCAT/EU-Interreg, duCAT (NL-STW), EURECA (EU-FP7), SeDI & CloudAtlas (EU-EUREKA), TruT (NL-CTMM), DLRA (NL-NVO)

"Data! Data! Data!" he cried impatiently.
"I can't make bricks without clay."
Sherlock Holmes
Data driven healthcare

Big Data in healthcare

- 2/3 petabytes in average hospital
  (500 years 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
by 2020
80% unstructured
data increased

The problem of Big Data – The doctor is drowning

- 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)


“We cannot predict outcomes of individual treatments”
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 investigation of listed clinical data sources rather than prospective randomized controlled clinical trials

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. Examples: EMLA, CAT (www.eurocat.info) ASCO’s CancerLinQ.

Data landscape

- Clinical research
  - patients
  - features
  - 285 data points
- Clinical registries
  - patients
  - features
  - 240 data points
- Clinical routine
  - patients
  - features
  - 2000 data points

Data Landscape – completing data
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 \(\rightarrow\) FAIRify medical databases

EuroCAT: distributed learning network

EuroCAT
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): cancerdata.org/protocols

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

- Findable: the data should be uniquely and persistently identifiable (PID) and other researchers should be able to find your data.
- Accessible: the conditions under which the data can be used should be clear to machines and humans.
- Interoperable: data should be machine-readable and use terminologies, vocabularies, or ontologies that are commonly used in the field.
- Reusable: 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 Open Science Cloud

BRINGING TOGETHER CURRENT AND FUTURE DATA INFRASTRUCTURES

A federated, open environment for sharing scientific data
Linking data
Connecting across borders and scientific disciplines

Open and seamless services to produce and reuse research data
Connecting scientists globally
Long term and sustainable
Improving science


---

Data as increasingly FAIR Digital Objects

Re-useless data (99%)

Findable

Data as FAIR metadata

A FAIR + open

B FAIR + closed

C FAIR + unknown

D Open data

E Open data

F Open data

Mons 2017, DOI:10.3233/ISU-170824
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. Bootstrapping 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 outliers 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

Acknowledgements

• Varian, Palo Alto, CA, USA
• Siemens, Malvern, PA, USA
• RTDG, Philadelphia, PA, USA
• MAASTRO, Maastricht, Netherlands
• Policlinico Gemelli, Rome, Italy
• UZ Ghent, Belgium
• Catharina Ziekenhuis, Eindhoven, Netherlands
• UZ Leuven, Belgium
• Radboud, Nijmegen, Netherlands
• University of Sydney, Australia
• UMCG, Groningen, Netherlands
• Liverpool and Macarthur CC, Australia
• CHU Liege, Belgium
• Umklinikum Aachen, Germany
• LGC Gent/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
Thank you for your attention

For more information:
www.europat.info
www.predictcancer.org
www.cancerdata.org
www.maastro.info
www.gridcsc.org
www.maastro.nl

FAIR adoption by European Commission
Model based approach

- Proton therapy introduction in the Netherlands
- Expensive and only 1800 slots
- ALARA vs protons for reduced toxicity
- Evidence based (e.g. paediatric) and model-based indications (HN, GBM, Lung, Brain, Prostate)

Widder et al. (iK) © 2017

Model based approach

Cheng et al. (iK) © 2017

There is an app for that

- PredictCancer.org
- Use today: Breast & ovarian
- Use tomorrow: Lung & prostate

There is an app for that

- PredictCancer.org
- Use today: Breast & ovarian
- Use tomorrow: Lung & prostate
Model-based proton therapy
Proton Decision Support (PRODECIS)
The example of healthcare efficiency

- Life expectancy vs. health expenditure, 1976 to 2013

---

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