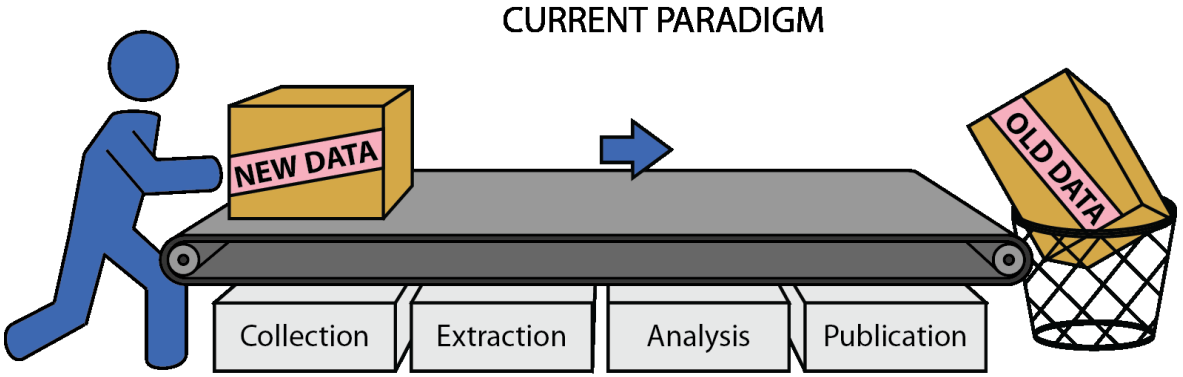
# Big data for health

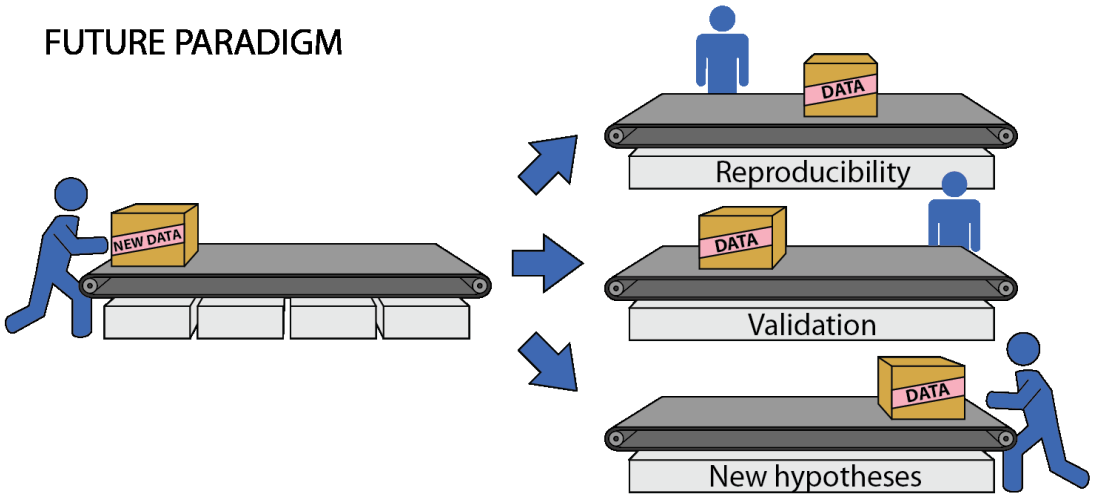
*Third pillars of CERN: "Storage and treatment of large amount of data and detailed simulation"*

Prof. Philippe Lambin

U.H. Maastricht

# Why? Reusing data



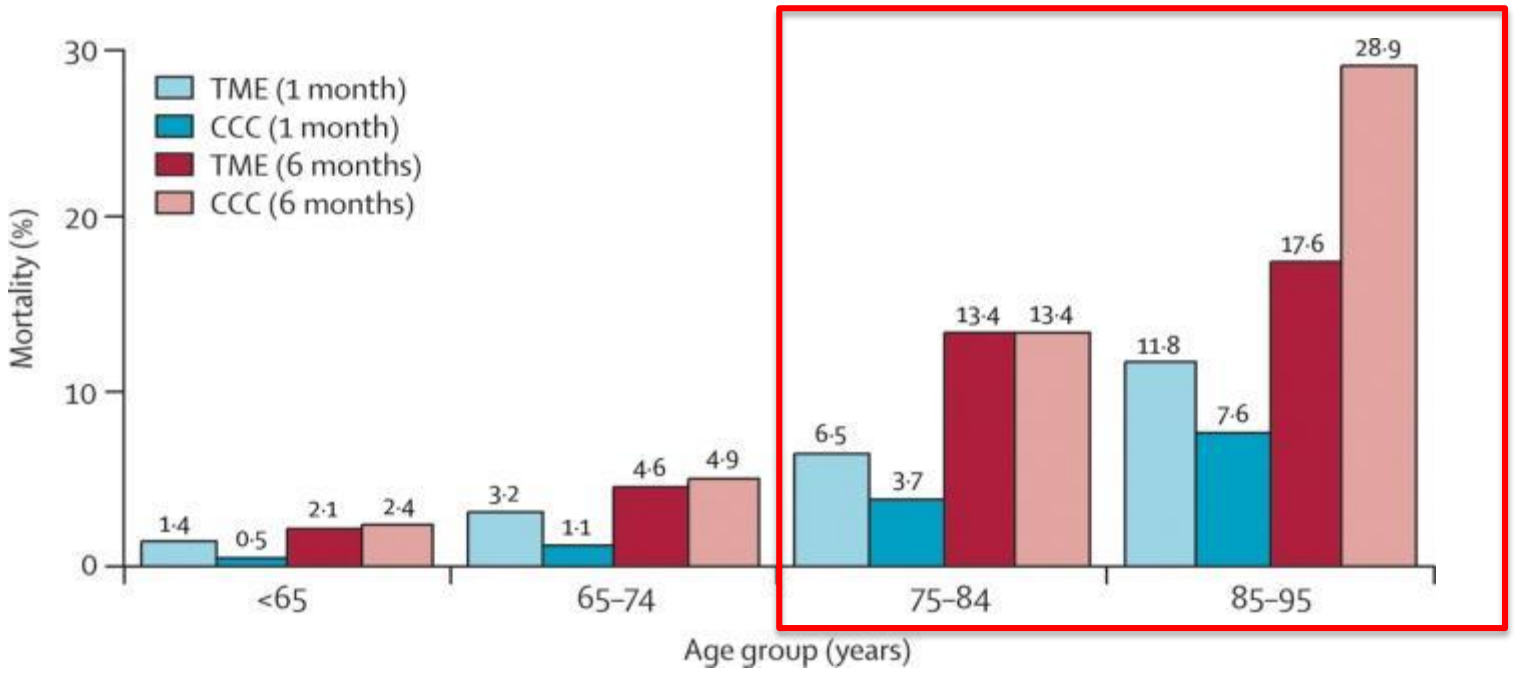




Modified from
Deasy et al.

# Why? Limitations of Evidence-based medicine


- Less then 3% of the patients
- Highly biased population
- Randomized trials rarely done for new technologies

# Example: having *no evidence* can have dramatic consequences




*Rutten et al. Lancet Oncology 2008; 9: 494*

| Conventional Clinical Research | Rapid Learning Health Care ("Big Data") |
| --- | --- |
| High data quality | Low data quality |
| Low data quantity | High data quantity |
| **Controlled**<br>○ Assigned patients<br>○ "EORTC-RTOG grade" QA/Protocol<br>○ Biobanking, translational research | **Reality**<br>○ Unassigned patients<br>○ "Clinical grade" QA/Protocol<br>○ Ad hoc biobanking/translational research |

Universiteit Maastricht

Relton C et al. BMJ. 2010; Burbach et al. Trials 2015; Lambin et al. Acta Oncol 2015

MAASTRO

# Example of clinically relevant questions

Treatment of

- 80 years old rectal cancer?
- 70 years old Stage IIIB NSCLC?
- 60 years old prostate cancer with oligometast
- Local relapse of a stage 3 oropharynx?
- Cervix cancer stage 3, HIV+
- …

Big Data to save lives

Watch the animation:
http://youtu.be/ZDJFOxpwqEA

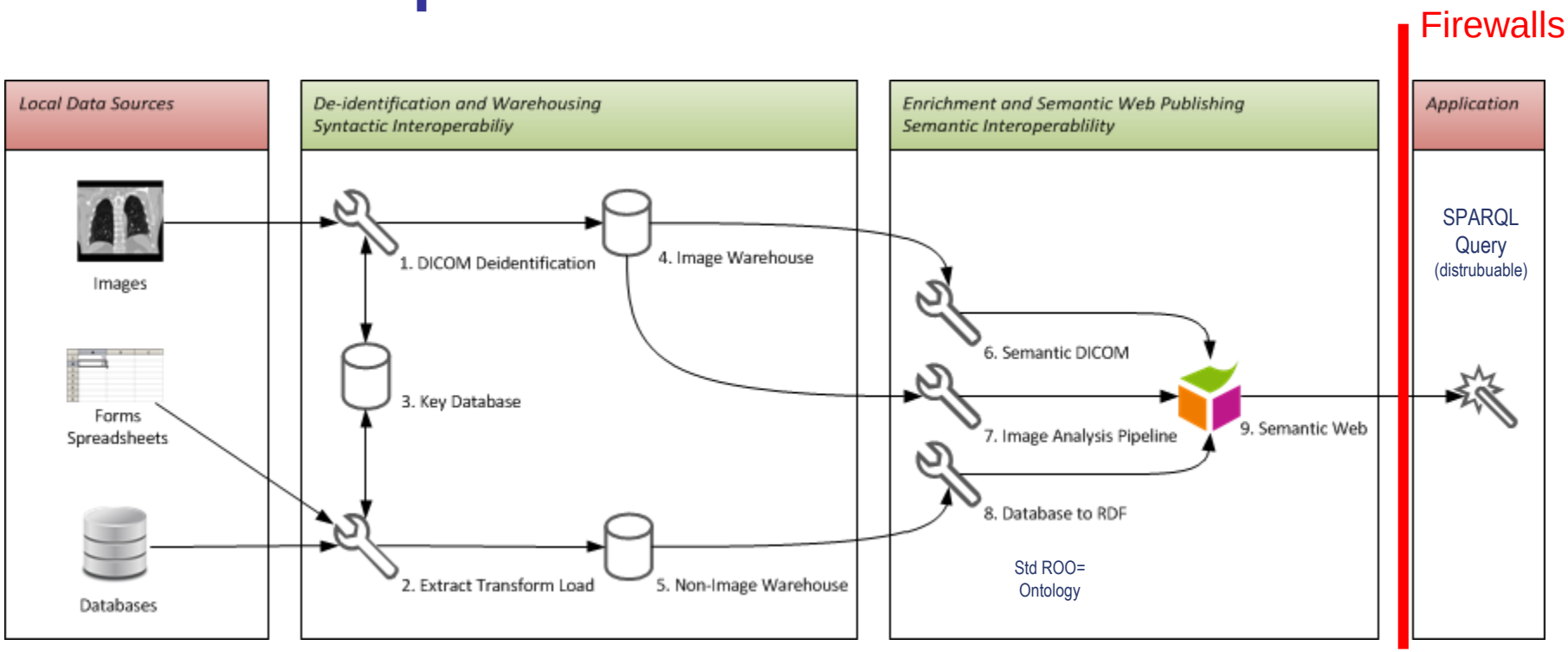
# ”Pan-omics approach”: Multifactorial Decision Support System


**But we need Data, preferably *all* of them**

**How?**

Lambin et al. Nature Rev. Clin. Oncol.

# In-hospital infra & de-identification

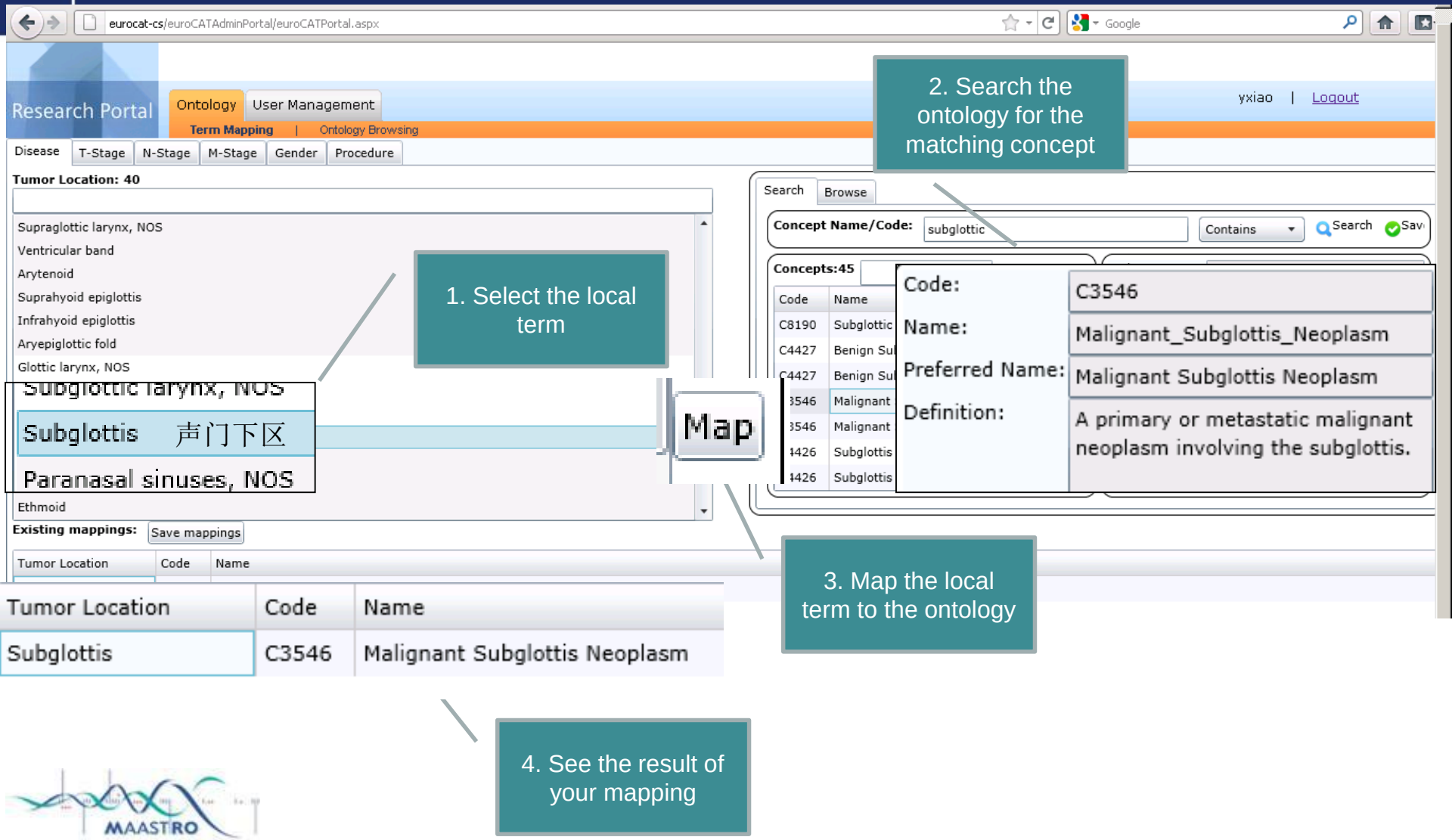



*SPARQL = "Simple Protocol And RDF Query Language"*
*a query language for databases*

Deidentification:

- Removal of obvious patient identifiers (name, MRN, social security number, email etc.)
- Assign a persistent token pseudonym
- Change (data banding) of obvious but required patient identifiers (everyone born and died on the 15th of the month, part of the postal code)
- No individual patient data leaves the hospital

# Ontology – International Coding System

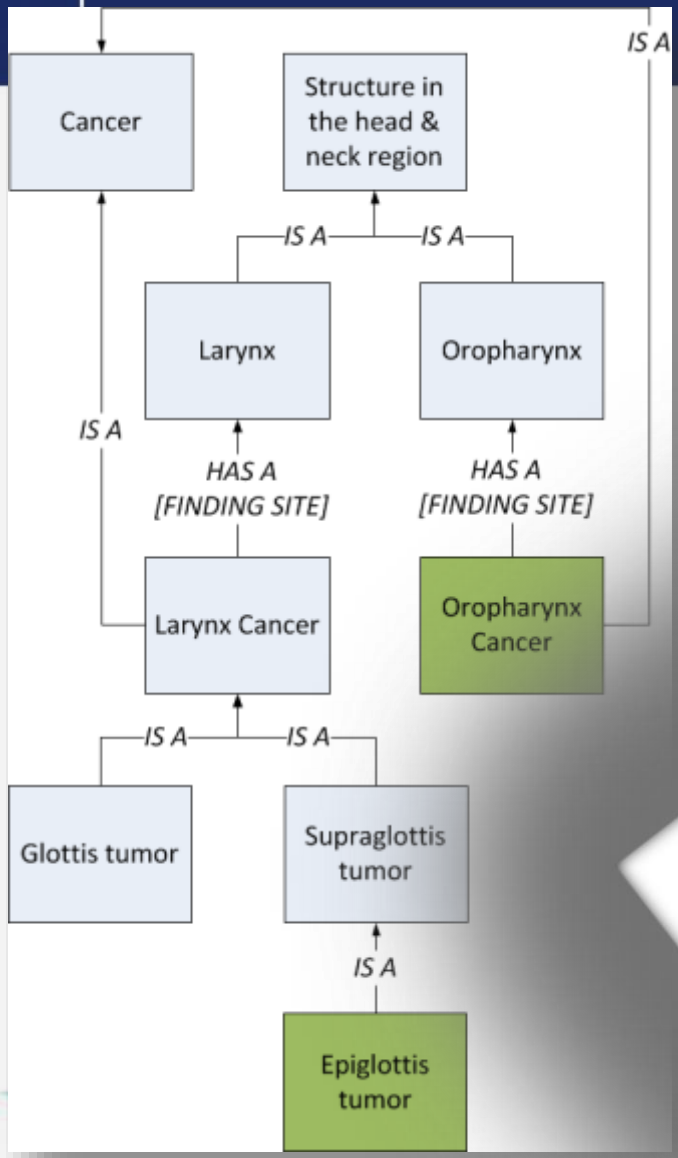


# The Semantic Web

- The **Semantic Web** is an extension of the Web through standards by the World Wide Web Consortium (W3C). The standards promote common data formats and exchange protocols on the Web.
- According to the W3C, "The Semantic Web provides a common framework* that allows data to be shared and reused across application, enterprise, and community boundaries". The term was coined by Tim Berners-Lee for a *web of data that can be processed by machines*.

*SPARQL is a semantic query language for databases

# An ontology is more than a dictionary

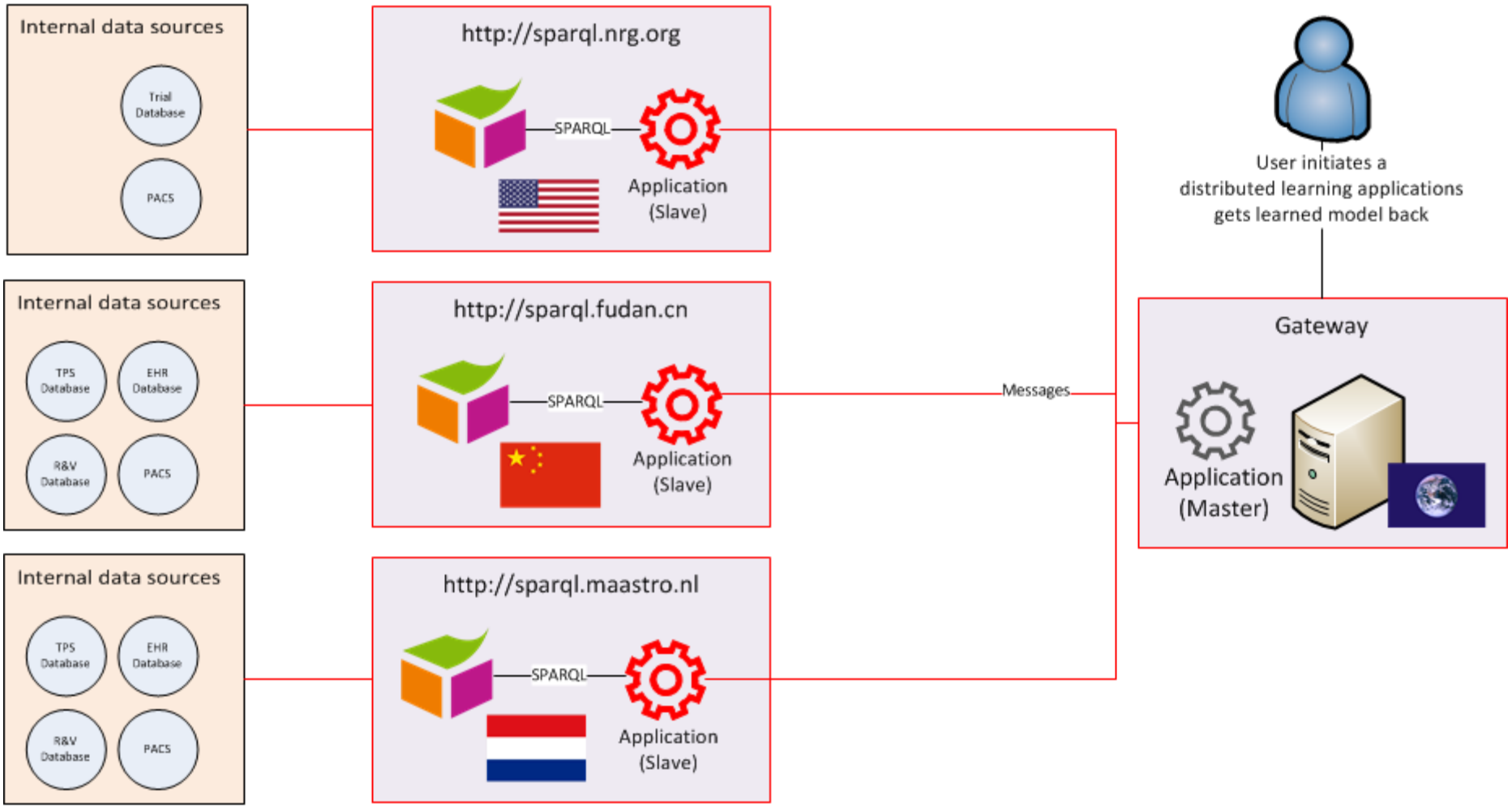



Ontology is a set terms & their **relationships**.

Then we have "machine readable data" accessible to Artificial Intelligence

## ***Semantic box*** *(the secondary research database*

## ***Hospitals***

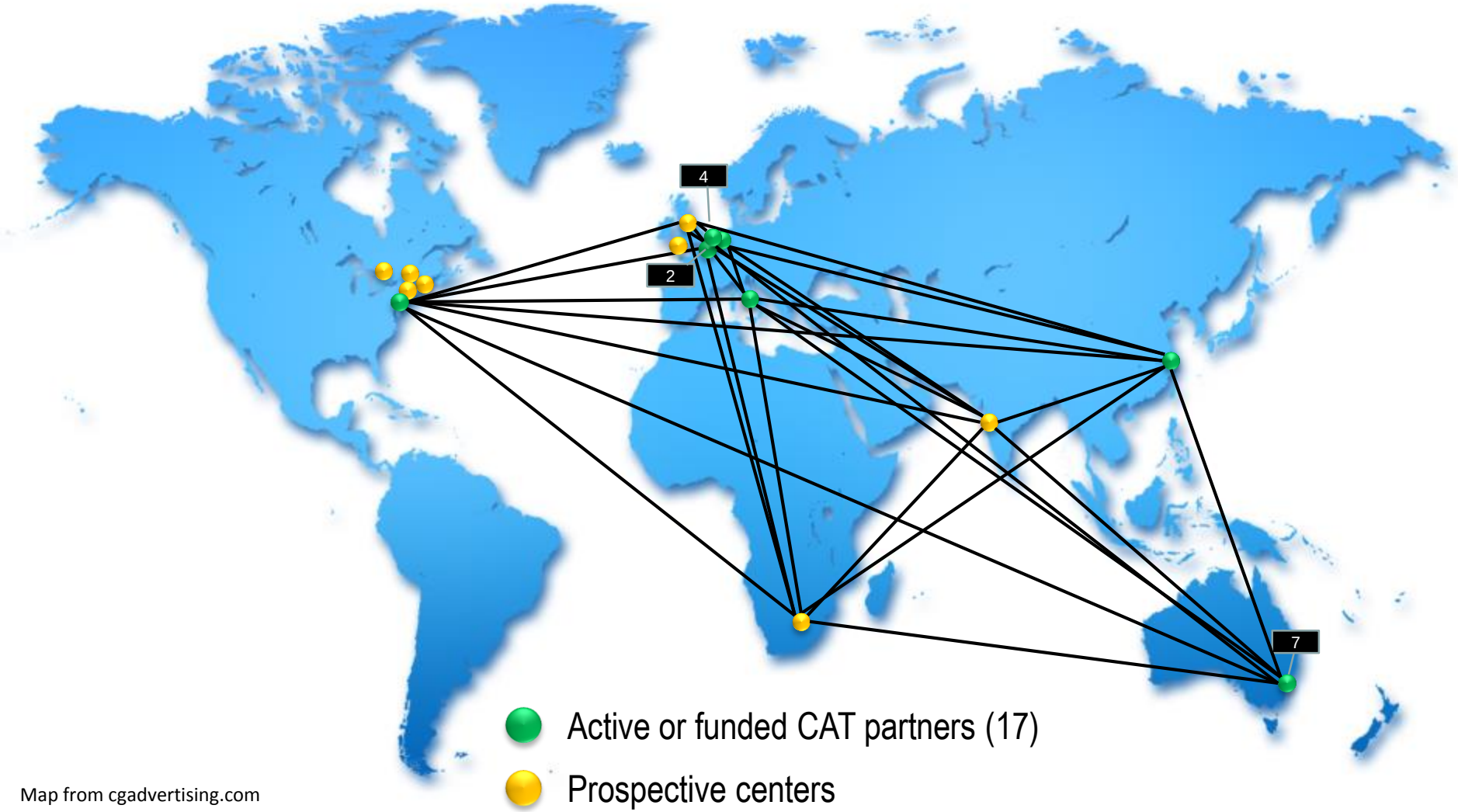



*SPARQL : Query language for application*

Universiteit Maastricht

MAASTRO

# Funded: euroCAT, duCAT, chinaCAT, VATE, ozCAT
# New: ukCAT, indiaCAT




Map from cgadvertising.com

# What next?: The patient managing its own data

Data =

**Our vision in 2 min:**

***"from hospital to patient"***

# From data to models to

# Virtual patient?

# The 5 P's of modern medicine

(modified from Leroy Hood)

« P » for Personalized

« P » for Preventive

« P » for Predictive

« P » for *Participatory*

« P » for Parcimonious

# Shared Decision Making 1.0 with Decision aids

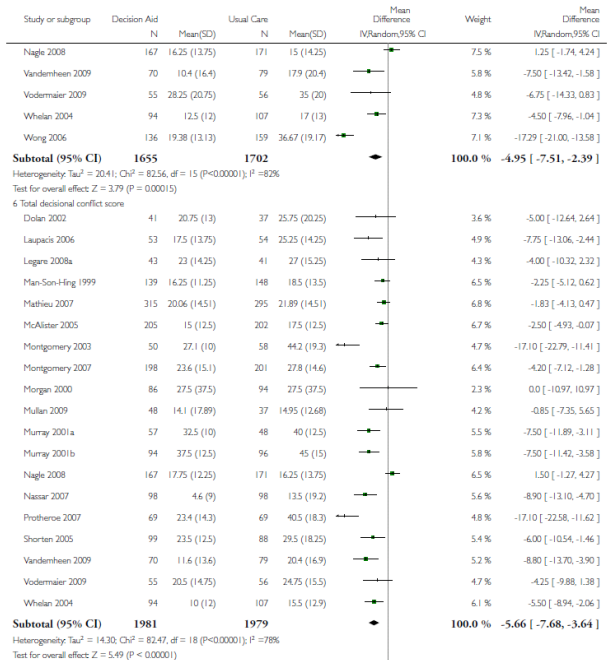
**Decision aids for people facing health treatment or screening decisions (Review)**

Stacey D, Bennett CL, Barry MJ, Col NF, Eden KB, Holmes-Rovner M, Llewellyn-Thomas H, Lyddiatt A, Légaré F, Thomson R

THE COCHRANE COLLABORATION®

This is a reprint of a Cochrane review, prepared and maintained by The Cochrane Collaboration and published in *The Cochrane Library* 2012, Issue 5

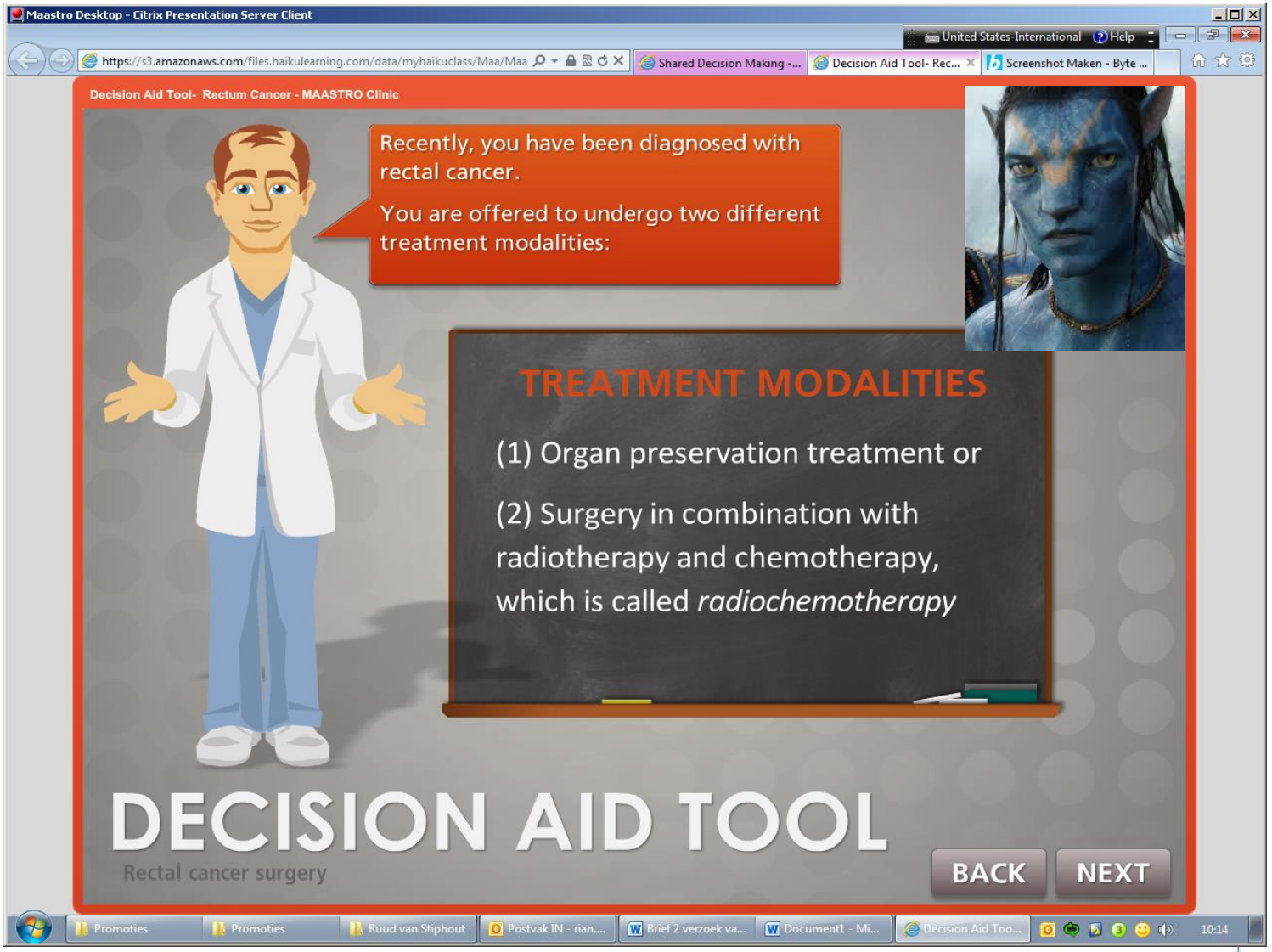
http://www.thecochranelibrary.com



| Study or subgroup | Decision Aid N | Mean(SD) | Usual Care N | Mean(SD) | Mean Difference IV,Random,95% CI | Weight | Mean Difference IV,Random,95% CI |
|---|---|---|---|---|---|---|---|
| Nagle 2008 | 167 | 16.25 (13.75) | 171 | 15 (14.25) | | 7.5 % | 1.25 [ -1.74, 4.24 ] |
| Vandemheen 2009 | 70 | 10.4 (16.4) | 79 | 17.9 (20.4) | | 5.8 % | -7.50 [ -13.42, -1.58 ] |
| Vodermaier 2009 | 55 | 28.25 (20.75) | 56 | 35 (20) | | 4.8 % | -6.75 [ -14.33, 0.83 ] |
| Whelan 2004 | 94 | 12.5 (12) | 107 | 17 (13) | | 7.3 % | -4.50 [ -7.96, -1.04 ] |
| Wong 2006 | 136 | 19.38 (13.13) | 159 | 36.67 (19.17) | | 7.1 % | -17.29 [ -21.00, -13.58 ] |
| **Subtotal (95% CI)** | **1655** | | **1702** | | | **100.0 %** | **-4.95 [ -7.51, -2.39 ]** |
| Heterogeneity: $Tau^2$ = 20.41; $Chi^2$ = 82.56, df = 15 (P<0.00001); $I^2$ =82% | | | | | | | |
| Test for overall effect: Z = 3.79 (P = 0.00015) | | | | | | | |
| 6 Total decisional conflict score | | | | | | | |
| Dolan 2002 | 41 | 20.75 (13) | 37 | 25.75 (20.25) | | 3.6 % | -5.00 [ -12.64, 2.64 ] |
| Laupacis 2006 | 53 | 17.5 (13.75) | 54 | 25.25 (14.25) | | 4.9 % | -7.75 [ -13.06, -2.44 ] |
| Legare 2008a | 43 | 23 (14.25) | 41 | 27 (15.25) | | 4.3 % | -4.00 [ -10.32, 2.32 ] |
| Man-Son-Hing 1999 | 139 | 16.25 (11.25) | 148 | 18.5 (13.5) | | 6.5 % | -2.25 [ -5.12, 0.62 ] |
| Mathieu 2007 | 315 | 20.06 (14.51) | 295 | 21.89 (14.51) | | 6.8 % | -1.83 [ -4.13, 0.47 ] |
| McAlister 2005 | 205 | 15 (12.5) | 202 | 17.5 (12.5) | | 6.7 % | -2.50 [ -4.93, -0.07 ] |
| Montgomery 2003 | 50 | 27.1 (10) | 58 | 44.2 (19.3) | | 4.7 % | -17.10 [ -22.79, -11.41 ] |
| Montgomery 2007 | 198 | 23.6 (15.1) | 201 | 27.8 (14.6) | | 6.4 % | -4.20 [ -7.12, -1.28 ] |
| Morgan 2000 | 86 | 27.5 (37.5) | 94 | 27.5 (37.5) | | 2.3 % | 0.0 [ -10.97, 10.97 ] |
| Mullan 2009 | 48 | 14.1 (17.89) | 37 | 14.95 (12.68) | | 4.2 % | -0.85 [ -7.35, 5.65 ] |
| Murray 2001a | 57 | 32.5 (10) | 48 | 40 (12.5) | | 5.5 % | -7.50 [ -11.89, -3.11 ] |
| Murray 2001b | 94 | 37.5 (12.5) | 96 | 45 (15) | | 5.8 % | -7.50 [ -11.42, -3.58 ] |
| Nagle 2008 | 167 | 17.75 (12.25) | 171 | 16.25 (13.75) | | 6.5 % | 1.50 [ -1.27, 4.27 ] |
| Nassar 2007 | 98 | 4.6 (9) | 98 | 13.5 (19.2) | | 5.6 % | -8.90 [ -13.10, -4.70 ] |
| Protheroe 2007 | 69 | 23.4 (14.3) | 69 | 40.5 (18.3) | | 4.8 % | -17.10 [ -22.58, -11.62 ] |
| Shorten 2005 | 99 | 23.5 (12.5) | 88 | 29.5 (18.25) | | 5.4 % | -6.00 [ -10.54, -1.46 ] |
| Vandemheen 2009 | 70 | 11.6 (13.6) | 79 | 20.4 (16.9) | | 5.2 % | -8.80 [ -13.70, -3.90 ] |
| Vodermaier 2009 | 55 | 20.5 (14.75) | 56 | 24.75 (15.5) | | 4.7 % | -4.25 [ -9.88, 1.38 ] |
| Whelan 2004 | 94 | 10 (12) | 107 | 15.5 (12.9) | | 6.1 % | -5.50 [ -8.94, -2.06 ] |
| **Subtotal (95% CI)** | **1981** | | **1979** | | | **100.0 %** | **-5.66 [ -7.68, -3.64 ]** |
| Heterogeneity: $Tau^2$ = 14.30; $Chi^2$ = 82.47, df = 18 (P<0.00001); $I^2$ =78% | | | | | | | |
| Test for overall effect: Z = 5.49 (P < 0.00001) | | | | | | | |

# Shared Decision Making 2.0: model-based virtual patient or

# *Avatar-based Shared Decision making*



# Data

## Patient avatar (or similar patients)

Game changer!

## Simulation/DSS:

Virtual treatments
Virtual clinical trials
Virtual scenarios with preventive personalized interventions
…

# Take home message: we need

1. Privacy-preserving Big data to build
2. Multifactorial decision support systems
3. And Shared decision making
4. Patient avatars (= model-based virtu
   patients or similar patients)
5. Used for *simulations* of virtual tr
   in virtual hospitals and virtu
   trials.

Big Data to save lives

# What next?

## Backcasting*

*= a planning method that starts with defining a desirable future and then works backwards to identify policies and programs that will connect the future to the present

# Thank you for your attention

1. Privacy-preserving Big data to build
2. Multifactorial decision support systems
3. And Shared decision making
4. Patient avatars (= model-based virtual patients or similar patients)
5. Used for *simulations* of virtual treatment in virtual hospitals and virtual clinical trials.

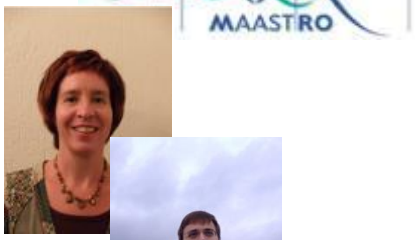
# Acknowledgements

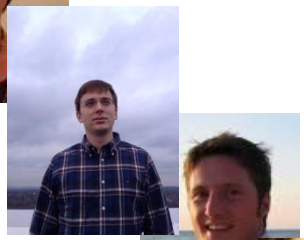

- Policlinico Gemelli, Roma, Italy
- UH Ghent, Belgium
- UH Leuven, Belgium
- UH Nijmegen, Netherlands
- …

- CHU Liege, Belgium
- Uniklinikum Aachen, Germany
- LOC Genk/Hasselt, Belgium
- Catherina Zkh Eindhoven, Netherlands


**Main MAASTRO collaborators**

- Andre Dekker
- Cary Oberije
- Timo Deist
- Erik Roelofs
- Arthur Jochems
- Sean Walsh
- Ralph Leijenaar
- Janita van Timmeren





# Reserve slides

# Open source data of publications: www.cancerdata.org

*Sharing medical data for cancer*

## ncerData

site is an effort of the Medical Informatics and Knowledge Engineering team of Maastro Clinic, Maastricht, The Netherlands. Our activities in the field of nalysis and data modelling are visible in a number of projects we are running. he Links for more information.

**riven**

**Follow us**

Follow @CancerData

**Navigation**

About

Universiteit Maastricht

# Protocols

By using "Big Data" we can address clinical problems. Analyzing the massive amount of clinical information that is available in digital format will make it possible to create a rapid learning health care system in which we develop, validate and update predictive tools to assist clinicians in personalizing treatment. Yet, some hurdles have to be taken. Besides technological, privacy and security issues, the most important bottleneck is the quality of the available clinical data.

To derive insights from data, it is critical that they are accurate and relatively complete. Thus, relevant variables should be collected and their definition should be clear. Also, machine learning algorithms require structured data while currently the richest source of clinical data, the clinicians' notes, is unstructured. However, writing research protocols is time-consuming and many clinicians lack time to do so, although they recognize the importance of collecting high quality data. We therefore created this open source research protocol repository. We anticipate that this initiative will stimulate centers to participate in outcomes research and will improve standardization and quality of data.

| Title | Last Update |
|---|---|
| Standard Follow Up Program For Head And Neck Cancer Patients | 2015-04-19 |
| EuroCAT Umbrella Protocol for NSCLC | 2015-04-16 |

update request

Enter your name here

Please add your name

**Affiliation** *

Please enter your affiliati

Please add your affiliation

**E-mail** *

name@domain.net

Please add your e-mail address so we can notify you of updated protocols.

Submit

# EuroCAT Umbrella Protocol for NSCLC

**Tags**: NSCLC, EuroCAT, protocol, data collection

For the EuroCAT project, a research protocol that describes a standardized data collection for non-small cell lung cancer was written and has been approved by the Medical Ethical Board of our hospital. A copy of the protocol and the appendices, including scoring of side effects, quality of life questionnaires and optional biobank procedure can be downloaded below. Patient information and the informed consent sheet are available in four languages (English, Spanish, French and Chinese).

It is allowed to adapt the documents, so that they match the requirements of your hospital and country. You can either collect data in your Electronic Medical Record System or use the eCRFs, that have already been created by us, and which are also freely available. It is also possible to publish your own "ready to use" protocol online and let other institutes participate in your research.

Please find all data below. If you leave your email address at the right of the screen, we can contact you if an updated version of the protocol is available.

| Attachment | Size |
|---|---|
| Material Transfer Agreement (doc) | 40.5 KB |
| EuroCAT Umbrella Protocol NSCLC (pdf) | 152.62 KB |
| Appendix A – Data Collection (pdf) | 29.24 KB |
| Appendix B – CTC Toxicity (pdf) | 13.69 KB |
| Appendix E – Timepoints (pdf) | 12.1 KB |
| Appendices – Chinese (zip) | 224.9 KB |
| Appendices – Dutch (zip) | 132.47 KB |

## Protocol update request

**Name** *

Enter your name here

Please add your name

**Affiliation** *

Please enter your affiliati

Please add your affiliation

**E-mail** *

name@domain.net

Please add your e-mail address so we can notify you of updated protocols.

Submit

# Can you give me

# examples

# of new knowledge coming from Big Data approaches

# The Radiomic hypothesis

## One can extract *more* quantitative information from standard imaging

Radiology:

- Implicit knowledge
- Interpretability

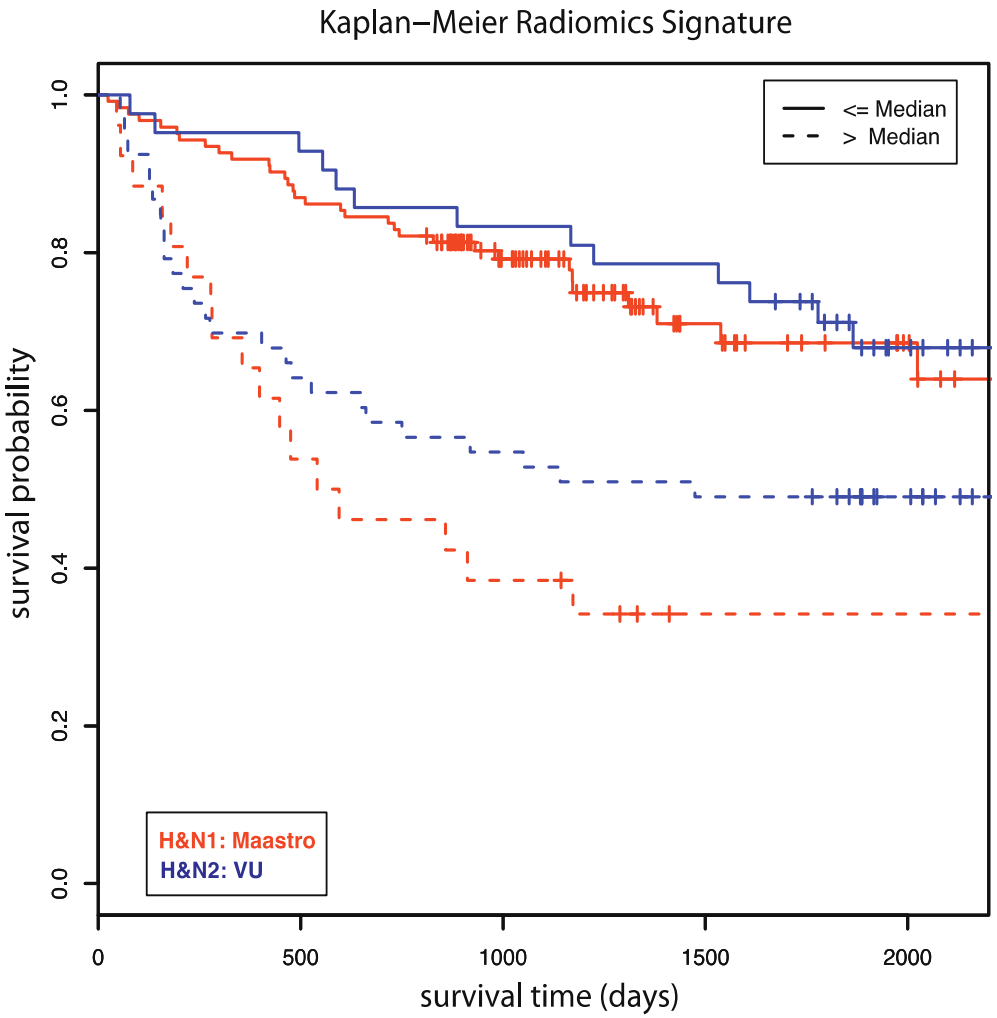
QUANTIFICATION

**RADIOMICS**
Extract *quantitative* features from images

Universiteit Maastricht

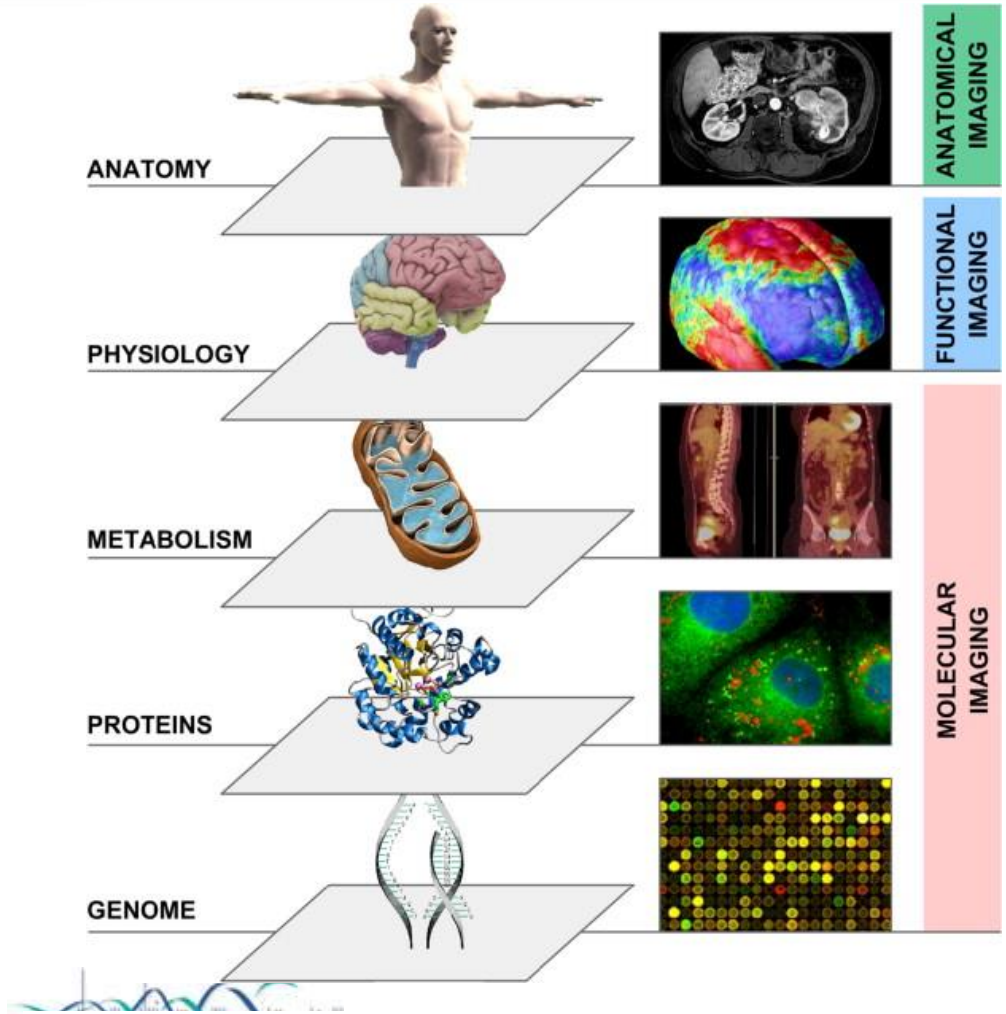
Lambin et al. EJC, 2012; Aerts, Lambin et al. Nature Commun 2014

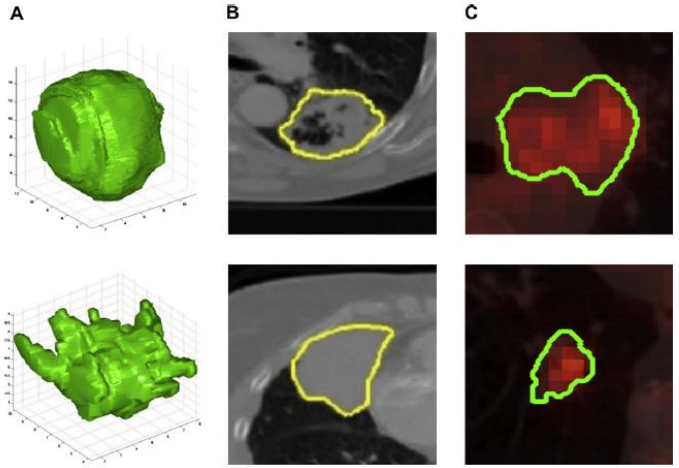
# Predict survival in Lung and Head & neck cancer better then TNM

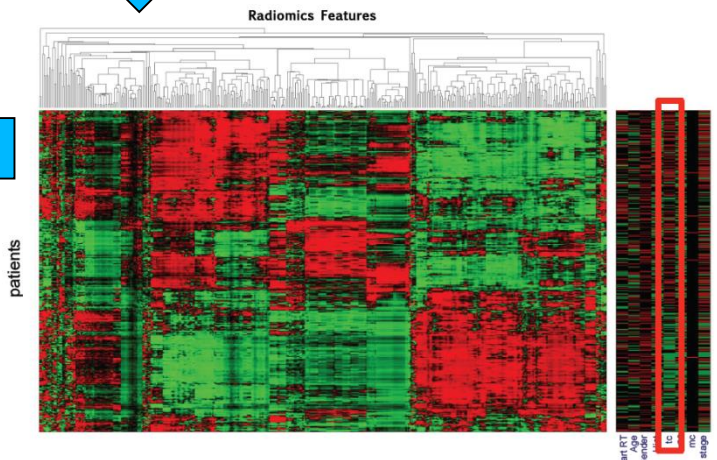



Aerts…Lambin, Nature Commun 2014; Leijenaar et al. Acta Oncol 2015

# Entering the OMICS era… Radiomics




Lambin et al. EJC, 2012; Aerts, Lambin et al. Nature Commun 2014

Watch the animation:
http://youtu.be/Tq980GEVP0Y

Or the website: www.radiomics.org

# Take home message

1. We need Decision Support Systems (DSS = a “meta TPS”) to manage the large quantity of data and implement Personalized medicine in radiotherapy in particular for protontherapy due to its costs.

2. Two complementary approaches: conventional clinical trials (+ data reuse) + “Big Data approach” (Rapid Learning Health Care).

3. Building cancer informatics tools to enable analysis, exploration, and rapid evaluation of novel therapies or stratification e.g. Distributed learning based on semantic web technology.

4. DSS facilitate Share Decision Making, participative precision medicine and cost effective Health care (the 4$^{th}$ & 5$^{th}$ “P”). One key example could be protontherapy.