Knowledge Engineering in Radiation Oncology –
A real-world application of Semantic Web technology

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
Disclosures

• Research collaborations incl. funding
  • Varian (VATE, chinaCAT, euroCAT), Siemens (euroCAT), Sohard (SeDI), Philips, Xerox (EURECA)

• Public research funding
  • Radiomics (USA-NIH/U01CA143062), euroCAT(EU-Interreg), duCAT (NL-STW), EURECA (EU-FP7), SeDI (EU-EUREKA), TraIT (NL-CTMM), DLRA (NL-NVRO)

• Spin-offs and commercial ventures
  • MAASTRO Innovations B.V. (CSO)
Contents

• Why did we start the CAT* project?
• Barriers to sharing data
• What are we doing with Semantic Web
• A note on Big Data
• Pilot project in Liverpool, NSW

• *CAT=euroCAT, duCAT, VATE, chinaCAT, SeDI etc.
Why we started the CAT project

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
Can we predict a tulip’s color by looking at the bulb?

http://www.amystewart.com
Predicting the tulip color

AUC

1.00

0.72

0.50
Experiment

AUC 1.00

AUC 0.72

AUC 0.50
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

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

J Clin Oncol 2010;28:4268
JMI 2012 Friedman, Rigby
Why?

- In radiotherapy we treat patients using technology.
- Getting RCT evidence on technology is hard.
- Already we can’t predict the outcome of RT patients.
- This will become worse as our knowledge of cancer and technology increases.
- If we can’t predict outcome, the treatment decisions will be suboptimal.

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 7133220; 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]
CAT is about improving medical decisions via the creation of an alternative evidence base:

Your Clinic

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
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:

Radiotherapy CAT ([www.eurocat.info](http://www.eurocat.info))

ASCO’s CancerLinQ
Build Decision Support Systems to individualize patient care by using machine learning to extract multifactorial personalized prediction models from existing databases containing all data on all patients that are validated in external datasets.
Classic linked data solution

• Sharing standardized, highly curated multi-domain data from clinical research programs
  • Via supplemental material / publications
  • Open-data
  • Via the Semantic Web
Back of the envelope

Link clinical radiotherapy treatment planning CT scans for Stage I-IIIB NSCLC to survival

• 10 years incidence 141M Cancer
• 18M Lung cancer
• 14M NSCLC
• 7M Stage I-IIIB
• 3.5M Stage I-IIIB RT

How can we get these 3.5M patients?

http://www.cancerresearchuk.org
Barriers to sharing data and a way to overcome these

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
Barriers to sharing data

[..] the problem is not really technical [...]. Rather, the problems are **ethical, political, and administrative**.

*Lancet Oncol* 2011;12:933

1. Administrative (time to capture, time to curate)
2. Political (value, authorship)
3. Ethical (privacy)
4. Technical
CAT 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
• The data has to be readable by an application (i.e. not a human)

• Solution: Sharing standardized highly curated research data
• Solution: Not-sharing non-standardized non-curated clinical data
Under the hood – Publish on the Semantic Web
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

1. Select the local term
2. Search the ontology for the matching concept
3. Map the local term to the ontology
4. See the result of your mapping
Funded: euroCAT, duCAT, chinaCAT, VATE
New: ozCAT, ukCAT, indiaCAT, “Big Machine” (Can.)
High level Infra

Internal data sources
- Trial Database
- PACS

Internal data sources
- TPS Database
- EHR Database
- R&V Database
- PACS

Internal data sources
- TPS Database
- EHR Database
- R&V Database
- PACS

User initiates a distributed learning application gets learned model back

Gateway

Application (Master)

Application (Slave)

Application (Slave)

Application (Slave)

http://sparql.fudan.cn

http://sparql.maastro.nl

http://sparql.nrg.org

SPARQL

SPARQL

SPARQL

SPARQL

© MAASTRO 2014
Distributed learning

Step 1: Sending parameters
- Master
- Center 1
- Center 2
- Center n

Step 2: Waiting
- Master
- Center 1
- Center 2
- Center n

Step 3: Waiting for all local models
- Master
- Center 1
- Center 2
- Center n

Step 4: Convergence criteria reached?
- No
- Master
- Center 1
- Center 2
- Center n

Step 5: Update parameters based on local models
- Master

Step 6: Create final model
- Master
Does all of that work? euroCAT’s example

- Distributed learning = Centralized learning
- Distributed learning better than learning on individual data

<table>
<thead>
<tr>
<th>Learn in</th>
<th>Validate in</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aachen (n=7)</td>
<td>Liège (n=186)</td>
<td>0.61</td>
</tr>
<tr>
<td>Eindhoven (n=32)</td>
<td>Liège (n=186)</td>
<td>0.72</td>
</tr>
<tr>
<td>Hasselt (n=45)</td>
<td>Liège (n=186)</td>
<td>0.68</td>
</tr>
<tr>
<td>Maastricht (n=52)</td>
<td>Liège (n=186)</td>
<td>0.75</td>
</tr>
<tr>
<td>Alle 4 samen (n=136)</td>
<td>Liège (n=186)</td>
<td>0.77</td>
</tr>
<tr>
<td>Alle 5 samen (n=322)</td>
<td>World (n=inf)</td>
<td>?</td>
</tr>
</tbody>
</table>

- 550 iterations, two hours (centralized < 1 min)
The computer curing cancer: Software is better than doctors at judging which treatments will work for patients. Computer models to help predicting how patients respond to certain cancer treatments are better than doctors. Scientists have discovered that computer models can outperform doctors at predicting the outcomes and responses of cancer patients to treatment. Mathematical models were more accurate than doctors at predicting how patients would respond to treatment. A study found that a computer model could predict cancer patients' responses to treatment better than doctors. The model is based on a database of 200,000 patients and has been shown to be more accurate in predicting responses to treatment.
A note on Big Data

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
Big Data in Medicine

Many people would have you believe this is due to genomics, but…

- **Volume:** “By 2015, the average hospital will have two-thirds of a petabyte of patient data, **80% of which will be unstructured image data** like CT scans and X-rays”.
- **Velocity:** “Medical Imaging archives are increasing by 20%-40% each year”
- **Variety:** 2D, 3D and 4D, modalities, contrast agents, multiple time points
- **Veracity:** Data quality and interpretation/annotation

The size of the genomics data from clinical patients at MAASTRO: 0 bytes.
Semantic DICOM?

- DICOM is the standard for medical imaging
- Digital Imaging and Communications in Medicine
- Paper standard, 20 parts and thousands of pages
- Syntactic standard
- A blog ("So you think HTML is hard? Try DICOM!")
  - The standard is long.
  - The standard is old.
  - The standard is ambiguous.
  - The standard is confusing.
  - Major vendors don’t comply to the standard.
Publishing image data as linked data
And how does all this benefit the patient?
An Australian Pilot

Andre Dekker, PhD
Medical Physicist
Information & Services Head
Knowledge Engineering Head
MAASTRO Clinic
Guideline vs. the real world in lung cancer

More survival, less quality of life

100%
50%
50%
Liverpool and Macarthur CC project

- DSS: Nomogram for OS for curatively treated NSCLC (AUC 0.75)
- Learned from curated data from MAASTRO
- Rapid Learning = real data
- “Commission” the DSS on real non-curated clinical data from Liverpool and Macarthur

- Is this possible with limited time & resources?
- Does the model work in non-curated data?
- What can we learn from it that is clinically relevant?
Patients found meeting inclusion criteria

- Practice insight: 50%+ of stage I-IIIB treated with non-radical treatments
- Sobering: 159 eligible from 3919 patients
## Missing data in included patients

<table>
<thead>
<tr>
<th>Site</th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAASTRO</td>
<td>322</td>
<td>159</td>
</tr>
<tr>
<td>Liverpool</td>
<td>159</td>
<td>322</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th># Patients</th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>322</td>
<td>159</td>
<td></td>
</tr>
</tbody>
</table>

### Age (years)

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>69 ± 10</td>
<td>68 ± 11</td>
</tr>
</tbody>
</table>

### Gender

<table>
<thead>
<tr>
<th></th>
<th>Male</th>
<th>Female</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training cohort</td>
<td>249 (77%)</td>
<td>73 (23%)</td>
</tr>
<tr>
<td>Commissioning cohort</td>
<td>113 (71%)</td>
<td>46 (29%)</td>
</tr>
</tbody>
</table>

### Stage

<table>
<thead>
<tr>
<th></th>
<th>I</th>
<th>II</th>
<th>IIIA</th>
<th>IIIB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training cohort</td>
<td>73 (23%)</td>
<td>29 (9%)</td>
<td>81 (25%)</td>
<td>134 (42%)</td>
</tr>
<tr>
<td>Commissioning cohort</td>
<td>26 (16%)</td>
<td>26 (16%)</td>
<td>62 (39%)</td>
<td>45 (28%)</td>
</tr>
</tbody>
</table>

### Missing

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>5 (2%)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

### ECOG

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>94 (29%)</td>
<td>49 (31%)</td>
</tr>
<tr>
<td>1</td>
<td>169 (53%)</td>
<td>72 (43%)</td>
</tr>
<tr>
<td>≥2</td>
<td>52 (16%)</td>
<td>13 (8%)</td>
</tr>
<tr>
<td>Missing</td>
<td>7 (2%)</td>
<td>25 (16%)</td>
</tr>
</tbody>
</table>

### FEV1 (%)

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>70 ± 24</td>
<td>77 ± 20</td>
</tr>
<tr>
<td>Missing</td>
<td>48 (15%)</td>
<td>95 (60%)</td>
</tr>
</tbody>
</table>

### Tumor load (cc)

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>106 ± 113</td>
<td>161 ± 147</td>
</tr>
<tr>
<td>p&lt;0.001</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Missing</td>
<td>36 (11%)</td>
<td>159 (100%)</td>
</tr>
</tbody>
</table>

### PLNS

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>143 (44%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>1</td>
<td>59 (18%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>2</td>
<td>44 (14%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>3</td>
<td>31 (10%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>≥4</td>
<td>30 (9%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>Missing</td>
<td>15 (5%)</td>
<td>159 (100%)</td>
</tr>
</tbody>
</table>

### 2 year OS

<table>
<thead>
<tr>
<th></th>
<th>Training cohort</th>
<th>Commissioning cohort</th>
</tr>
</thead>
<tbody>
<tr>
<td>Yes</td>
<td>103 (32%)</td>
<td>58 (36%)</td>
</tr>
<tr>
<td>No</td>
<td>219 (68%)</td>
<td>101 (64%)</td>
</tr>
<tr>
<td>Missing</td>
<td>0 (0%)</td>
<td>0* (0%)</td>
</tr>
</tbody>
</table>

### Still a lot of missing data -> imputation necessary (Bayesian Network)
Results

DSS works, but only to discriminate between good and medium/poor
Better than TNM stage
What did Liverpool learn?  What did MAASTRO learn?

- Rethink palliative treatments in good prognosis patients
- Rethink curative treatments in poor prognosis patients

Rapid learning: Expected survival gain with curative dose from 18 to ~60% in good prognosis patients

Rapid learning: No survival gain with curative dose in poor prognosis patients

routine data, realistic quality, good evidence?
Conclusion

- Link to youtube movie
- Semantic Web technology is helping us to open up clinics for rapid learning

Working on now

- Shape expressions
- Secure SPARQL
- Ontology extensions (esp. predicates) and integrations
- Cataloguing
- Physician facing interfaces for querying and analysis
- Probabilistic predicates – Bayesian Networks
Models built & validated

Lung cancer
- Survival
- Lung dyspnea
- Lung dysphagia

Rectal cancer
- Tumor response
- Local recurrences
- Distant metastases
- Overall survival

H&N cancer
- Local recurrences
- Overall survival
- Cost Effectiveness IMPT vs. IMRT

www.predictcancer.org
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
- 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

More info on:  
www.predictcancer.org  
www.eurocat.info  
www.cancerdata.org  
www.mistir.info
Thank you for your attention

More info on:
www.eurocat.info
www.predictcancer.org
www.cancerdata.org
www.mistir.info
www.maastro.nl