Rapid Learning Systems to improve patient outcomes & control health costs

Bharat R. Rao, PhD
Senior Director,
Innovations Center
Siemens Health Services

*This information about these product is preliminary. The product is under development and not commercially available for sale in the U.S. and its future availability cannot be ensured.

© Siemens 2012. All rights reserved.
Outline

- The Healthcare Problem
- Knowledge Solutions
- Rapid Learning Systems: Healthcare
- Focus on patient privacy
- Conclusions
The Healthcare Problem

**DATA OVERLOAD**
- Vast amounts of digital patient data available
- Images, Labs, Text, Demographics, Patient risk factors, …
- Genomics, Proteomics will worsen data overload

**THERAPY OVERLOAD**
- Many promising therapies.
- 1000s of drugs with genetic indications in FDA pipeline
- Hard to implement @ poc

**KNOWLEDGE OVERLOAD**
- Explosion in the amount of medical knowledge
- Estimated to double every 18-36 months

1. Healthcare costs growing explosively
2. Without matching improvement in the quality of care & patient outcomes
Individualizing care to every patient will:

Reduce Costs
- Cost of side effects
- Continued care costs
- Cost of ineffective treatment

Improve Quality
- Reduce side effects
- Select the best therapy for each patient

Patients can respond differently to the same medicine:

<table>
<thead>
<tr>
<th>Medicine Type</th>
<th>Percentage Effective on Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anti-Depressives</td>
<td>62%</td>
</tr>
<tr>
<td>Asthma Drugs</td>
<td>60%</td>
</tr>
<tr>
<td>Diabetes Drugs</td>
<td>57%</td>
</tr>
<tr>
<td>Arthritis Drugs</td>
<td>50%</td>
</tr>
<tr>
<td>Alzheimer’s Drugs</td>
<td>30%</td>
</tr>
<tr>
<td>Cancer Drugs</td>
<td>25%</td>
</tr>
</tbody>
</table>

Source: Brian B. Spear, Margo Heath-Chiozzi, Jeffrey Huff, “Clinical Trends in Molecular Medicine – Volume 7, Issue 5, 1st May 2001; Pages 201-204
A Healthcare Solution: Knowledge-based Medicine

Key role for data mining

DATA
- Vast amounts of digital patient data available
- Images, Labs, Text, Demographics, Patient risk factors, …
- Genomics, Proteomics will worsen data overload

THERAPY
- Many promising therapies.
- 1000s of drugs with genetic indications in FDA pipeline
- Hard to implement @ poc

KNOWLEDGE
- Explosion in the amount of medical knowledge
- Estimated to double every 18-36 months

Knowledge Utilization
Knowledge Discovery
Two aspects of Data Mining to drive Knowledge-Based Medicine

- Knowledge Utilization
  - Knowledge-driven decision support
  - Data aggregation from disparate sources
  - Inject knowledge (e.g., guidelines, mined models) via Probabilistic inference
  - Present conclusive findings

- Knowledge Discovery
  - Combine clinical, lab, pharma, text, imaging & genetic information
  - Discover predictive models
  - ... from multi-institution data
  - Enable increasing personalization of care
REMIND Knowledge Platform*: Architecture

Reliable Extraction & Meaningful Inference from Nonstructured Data

*Not considered a Medical device.
Outline

- The Healthcare Problem
- Knowledge Solutions
  - Computer Aided Diagnosis
  - Automated Quality Measurement
  - Individualized Medicine
- Rapid Learning Systems: Healthcare
- Focus on patient privacy
- Conclusions
Knowledge Solutions @ Siemens Healthcare

What are the different product lines?

| DETECT | Computer Aided Diagnosis:  
|        | Task: Analyze medical Images to identify abnormalities  
|        | Products: Classifier solutions for detection of cancers (lung, colon, breast), and cardiac abnormalities (PE, wall motion) deployed on multiple imaging modalities (X-Ray, CT, MR, ultrasound) at thousands of hospitals worldwide  
|        | Clinical approval (including FDA trial for LungCAD) in US & worldwide |
| MEASURE / IMPROVE | Quality Measurement & Improvement  
| | Task: Automatically measure adherence to quality guidelines and reduce cycle time for Quality Improvement  
| | Products / Cloud Services: Semi-automated EMR mining solution sold to ~300 US hospitals. Helps them qualify for ARRA meaningful use incentives. |
| PREDICT | Individualized Medicine (Personalized Medicine)  
| | Task: Mine all available patient data for risk prediction, therapy selection and patient prognosis  
| | Many research projects to develop predictive models, infrastructure.  
| | Example: 5-nation Euro-US rapid learning network to share data and learn personalized therapy selection models for lung cancer |

### Knowledge Solutions @ Siemens Healthcare

#### Research Challenges?

<table>
<thead>
<tr>
<th>DETECT</th>
<th>Computer Aided Diagnosis:</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>The breakdown of assumptions</td>
</tr>
<tr>
<td>2.</td>
<td>Highly unbalanced data</td>
</tr>
<tr>
<td>3.</td>
<td>Feature computation cost</td>
</tr>
<tr>
<td>4.</td>
<td>Incorporating domain knowledge</td>
</tr>
<tr>
<td>5.</td>
<td>No objective ground truth</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>MEASURE / IMPROVE</th>
<th>Quality Measurement &amp; Improvement</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>Access data from multi-source, multi-vendor systems without standards</td>
</tr>
<tr>
<td>2.</td>
<td>Combine structured and unstructured (text, images) data</td>
</tr>
<tr>
<td>3.</td>
<td>Integrate knowledge (guidelines) into machine learning</td>
</tr>
<tr>
<td>4.</td>
<td>Knowledge Maintenance (Handle changes in guidelines)</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>PREDICT</th>
<th>Individualized Medicine</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.</td>
<td>All the above plus</td>
</tr>
<tr>
<td>2.</td>
<td>Discover new knowledge (comparative effectiveness)</td>
</tr>
<tr>
<td>3.</td>
<td>Variation of data capture / tests available / preparation / guidelines across institutions</td>
</tr>
<tr>
<td>4.</td>
<td>Incorporate genomic information*</td>
</tr>
<tr>
<td>5.</td>
<td>Patient privacy (across institutions / across nations)</td>
</tr>
</tbody>
</table>

---

Outline

- The Healthcare Problem
- Knowledge Solutions
- Rapid Learning Systems: Healthcare
  - The promise of individualized medicine
  - Infrastructure
  - Machine learning
- Focus on patient privacy
- Conclusions
Example: Standard of Care Workflow for Rectum Cancer

- Today: Chemo-Radiotherapy (CRT) regimen followed by Rectal Surgery
- Rectal surgery is painful, with many side-effects
  - Morbidity: Colostomy, impotence, incontinence, plus surgery side-effects
  - Mortality

Pathologic Complete Response (pCR)
Upon resection after surgery, a significant number of tumors have a pCR, indicating that surgery may not have been needed.
Individualized Therapy workflow for Rectal Cancer*
Combine pre- and post-CRT to personalize therapy

Identify the sub-population of patients for whom alternate therapy should be considered

Diagnosis of Rectal Cancer → Pre-CRT Imaging + biomarkers → CRT: Chemo-Radio therapy → Surgery Decision Support

- Prob (pCR) = high
- else

Surgery → Rectal Surgery

Imaging follow-up / Alternate therapy → 2-year survival

Accurate Prediction of Survival & Side-effects (& Costs) helpful for rational decision-making

*This information about this product is preliminary. The product is under development and not commercially available for sale in the U.S. and its future availability cannot be ensured.

How well do doctors predict outcomes?

**Predict survival / sideeffects for NSCLC patients**

### Non Small Cell Lung Cancer

#### Survival
- 2 year survival
- 30 patients
- 8 MDs
- Retrospective
- **AUC: 0.57**

#### Side Effects
- Esophagitis
- 138 patients
- Prospective
- **AUC 0.53**

---

How good is Survival prediction in NSCLC?  
*Standard staging vs. Rapid Learning System*

AUC 0.75  
Stage IIIA 10 (14%)  
Stage IIIB 13 (19%)  
T4 12 (17%)

C Dehing-Oberije, S Yu, D De Ruysscher, S Meersschout, K Van Beek, Y Lievens, J Van Meerbeeck, W De Neve, B Rao, H van der Weide, P Lambin, “Development and external validation of prognostic model for 2-year survival of non-small-cell lung cancer patients treated with
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

- Traditionally, medicine advances via clinical studies: evidence-based medicine
- In the future, medicine could advance via rapid learning systems that learn from EMRs: evidence-generated medicine
EuroCAT: A Rapid Learning System for Personalizing cancer care for NSCLC

- Objectives: Clinical data sharing to facilitate development of models that can predict patient survival and side effects as a function of all patient data and therapy.
- Clinical trials for validation of model based personalized therapy selection
- Research collaboration with 5 hospitals across Netherlands, Belgium and Germany.
- Funded by grant from the InterReg region of EU
- More sites expected to join in the future.
- Siemens provides software to facilitate data capture & sharing across network

*This information about this product is preliminary. The product is under development and not commercially available for sale in the U.S. and its future availability cannot be ensured.*
Two aspects for this Rapid Learning System

- Theme 1: Build data sharing infrastructures
  - Integrated: euroCAT / duCAT / transCAT / Rome / EURECA
  - Imaging: Radiomics / QIN / caBIG/ TraIT
  - Genomics
  - Preserve Patient Privacy

- Theme 2: Machine learning for prediction modeling
  - Privacy Preserving Data Mining
Network 01/2012

- Active or funded CAT partners (10)
- Prospective centers (4)

Map from cgadvertising.com
Outline

- The Healthcare Problem
- Knowledge Solutions
- Rapid Learning Systems: Healthcare
- Focus on patient privacy
  - Privacy preserving via random projections
  - Dealing with different data in each institution
  - Distributed learning
- Conclusions
Focus: 3 aspects of research

1. Privacy preserving via random projections
2. Dealing with different data in each institution
3. Distributed learning
A real scenario: Description of the Patient data

- Patients form three institutions:
  - 455 inoperable NSCLC patients, stage I-IIIB referred to MAASTRO clinic (Netherlands) to be treated with curative intent.
  - 112 Patients from Gent hospital (Belgium)
  - 40 patients from Leuven hospital (Belgium), were also collected for this study.
- Data collected Between May 2002 and January 2007
- Same set of clinical variables for all patients in all centers were measured:
  - Gender
  - WHO performance status
  - Lung function prior to treatment (forced expiratory volume)
  - Number of positive lymph node stations
  - GTV
  - Dose corrected by time.

How do we use data from different institutions while preserving patient privacy?
Example: Cox Regression for Survival Analysis

- **Cox regression**, or the **Cox propositional-hazards model**, is one of the most popular algorithms for survival analysis
  - Defines a semi-parametric model to directly relate the predictive variables with the real outcome (i.e. the survival time)
  - Assumes a linear model for the log-hazard:
    \[
    \lambda(t | x_i) = \lambda_0(t) \exp[w^T x_i]
    \]

- We propose privacy-preserving Cox regression (PPCox) which is based on random projection
  - Provides accurate classification
  - Does not reveal private information
    \[
    \lambda_{HPP_Cox}(t | x_i) = \lambda_0(t) \exp[w^T B^T x_i]
    \]
How does this differ from commonly used random projection approach?

- **Relative distance preservation**: Finds a projection that is optimal at preserving the properties of the data that are important for the specific (learning) problem at hand.

- **Lower dimensionality in the projected space**: Produces an sparse mapping to generate a PP mapping that projects the data to a lower dimensional

- **Lower dimensionality in the input space**: Provides explicit feature selection as it also depends on a smaller subset of the original features

Privacy Preserving Cox Regression

How to choose the mapping matrix?

[ Yu et al, 2008 ]
Developing Privacy Preserving (PP) models for Survival

- We are using real data from several institutions across Europe to build models for survival prediction for non-small-cell lung cancer patients while addressing the potential privacy preserving issues.

Example showing the improvement in performance of a model trained using all the data available from multiple sites against models learned only using local available data.

Focus: 3 aspects of research

1. Privacy preserving via random projections
2. Dealing with different data in each institution
3. Distributed learning


S Yu, C Dehing-Oberije, D De Ruysscher, K Van Beek, Y Lievens, J Van Meerbeeck, W De Neve, G Fung, B Rao, P Lambin “Development, external validation and further improvement of a prediction model for survival of non-small cell lung cancer patients treated with (Chemo) radiotherapy,” International Journal of Radiation Oncology* Biology* Physics 72 (1), S60-S60

© Siemens 2012. All rights reserved.
How to learn better?

- AUC of 0.85 is minimal
- Survival model AUC 0.75
- More patients
- More variables
- Diversity

But....
As we increase the # of variables the variation of data stored at hospitals grows non-linearly
More variables from a simple CT
Knowledge Discovery: Learning Combined Diagnostics
Does incorporating more data help? (Collaboration with MAASTRO)

LOO ROC Plot for S2y (82pts, P/N: 24/58)

Combining clinical data from disparate sources improves prediction accuracy

Focus: 3 aspects of research

1. Privacy preserving via random projections
2. Dealing with different data in each institution
3. Distributed learning

Distributed Learning Architecture

Final Model Created

Central Server

Update Model

Model Server RTOG

Send Model Parameters

Model Server MAASTRO

Send Model Parameters

Model Server Eindhoven

Send Model Parameters

Learn Model from Local Data

Learn Model from Local Data

Learn Model from Local Data

Learn Model from Local Data

Send Average Consensus Model

Send Average Consensus Model

Send Average Consensus Model

Only aggregate data is exchanged between the Central Server and the local Servers.
Focus: 3 aspects of research

1. Privacy preserving via random projections
2. Dealing with different data in each institution
3. Distributed learning

- Putting it all together
Example: Standard of Care Workflow for Rectum Cancer

- Today: Chemo-Radiotherapy (CRT) regimen followed by Rectal Surgery
- Rectal surgery is painful, with many side-effects
  - Morbidity: Colostomy, impotence, incontinence, plus surgery side-effects
  - Mortality

Unfortunately, pCR is often hard to predict from pre-CRT data alone or detected pre-surgery via an imaging scan.

Pathologic Complete Response (pCR)
Upon resection after surgery, a significant number of tumors have a pCR, indicating that surgery may not have been needed.
Personalized Therapy workflow for Rectal Cancer*

Combine pre- and post-CRT to personalize therapy

**Identify the sub-population of patients for whom alternate therapy should be considered**

- **Diagnosis of Rectal Cancer**
  - Pre-CRT Imaging + biomarkers
  - CRT: Chemo-Radio therapy
  - Post-CRT Imaging + biomarkers

**Personalized Knowledge (IDS) for Decision Support**

By comparing patient imaging / biomarkers pre-CRT with patient imaging / biomarkers post-CRT it is possible to predict pathologic complete response with high accuracy.

- **Surgery Decision Support**
  - *Prob (pCR) = high*
  - *else*
  - Rectal Surgery

- **Imaging follow-up / Alternate therapy**
  - 2-year survival

*This information about this product is preliminary. The product is under development and not commercially available for sale in the U.S. and its future availability cannot be ensured.

Experimental Results
Post-CRT Dual PET vs Pretreatment PET only

- Post-CRT Dual PET validation (21 pts, 19% pCR)
- Post-CRT Dual PET (78 pts, 27% pCR)
- Pretreatment PET (85 pts, 16% pCR)
- Pretreatment NO PET validation (118 pts, 15% pCR)
- Pretreatment NO PET (768 pts, 19% pCR)

Creating Medical Knowledge
Mining large patient databases from multiple institutions

Data → REMIND Platform → Predictive Models

Knowledge
- Existing guidelines for cancer therapy
- Learned knowledge - Predictive models discovered from patient data
Available:
All published MAASTRO models (Lung, rectum, H&N)

Online input of patient data

Online calculation of probability of outcome and risk group stratification

Interpretation: If there would be a group of 100 patients with the same characteristics as this individual patient, 17 patients would be alive 2 years after the radiotherapy treatment. Due to the fact that a model can never be completely the same as the "real world" the number 17 could be lower or higher, but 17 is the most likely value.
Outline

- The Healthcare Problem
- Knowledge Solutions
- Rapid Learning Systems: Healthcare
- Focus on patient privacy
- Conclusions
[..] the problem is not really technical […]. Rather, the problems are ethical, political, and administrative. Lancet Oncol 2011;12:933

1. Administrative (time)
2. Political (value, authorship)
3. Ethical (privacy)

4. Technical
Why is this so critical now?

*Overload of data*

- Explosion of data
- Explosion of decisions
- Explosion of ‘evidence’*

*2010: 1574 & 1354 articles on lung cancer & radiotherapy = 7.5 per day
A Healthcare Solution: Knowledge-based Medicine

Key role for data mining

DATA
- Vast amounts of digital patient data available
- Images, Labs, Text, Demographics, Patient risk factors, …
- Genomics, Proteomics will worsen data overload

THROUGH
- Many promising therapies.
- 1000s of drugs with genetic indications in FDA pipeline
- Hard to implement @ poc

KNOWLEDGE
- Explosion in the amount of medical knowledge
- Estimated to double every 18-36 months
It will become *unethical* to ask Doctors to make, on their own, complex decisions.

We need a validated “decision support system”

…based *on all* the available data
Acknowledgements

- MAASTRO, Maastricht, Netherlands
- RTOG, Philadelphia, PA, USA
- Policlinico Gemelli, Roma, Italy
- UH Ghent, Belgium
- Catherina Zkh Eindhoven, Netherlands
- UZ Leuven, Belgium
- CHU Liege, Belgium

- Uniklinikum Aachen, Germany
- LOC Genk/Hasselt, Belgium
- Princess Margaret Hospital, Toronto, Canada
- The Christie, Manchester, UK
- UH Leuven, Belgium
- State Hospital, Rovigo, Italy

- Colleagues at the Knowledge Solutions Group @ Siemens Healthcare

More info on:  
www.predictcancer.org
www.eurocat.info
www.cancerdata.org
www.mistir.info
To wrest from nature the secrets which have perplexed philosophers in all ages, to track to their sources the causes of disease, to correlate the vast stores of knowledge, that they may be quickly available for the prevention and cure of disease—these are our ambitions.

Sir William Osler, 1849–1919

Father of Modern Medicine