Understanding the Interaction between Diseases using Big Healthcare Data

Peter Lucas  Martijn Lappenschaar  Arjen Hommersom

Radboud University Nijmegen
Institute for Computing and Information Sciences

12th December, 2014
Background

- Research focus: we try to combine knowledge and innovation from computing science with that of medicine.
- Clinical knowledge is not shallow and therefore requires a decent knowledge representation method (difficult issue, much progress during last three decades, but still a long way to go).
- Uncertainty is an essential ingredient of any form of clinical decision making.
- ⇒ Machine learning should take knowledge representation and uncertainty into account (compare medical statistics).
Rob is 67 years old and has been ill for some time, in particular he is currently treated for:

- diabetes mellitus type 2
- status after myocardial infarction
- chronic obstructive pulmonary disease

Rob is not unique ... 

- 2/3rd of patients older than 65 years have 2 or more disorders at the same time
- However, medicine is organised around single disorders! = problem of multimorbidity
Challenge: dealing with diseases and their interactions

- **Complexity**: many individual diseases and classes of disease
- **Probabilistic relationships**: uncertain interactions between diseases, regional, social, and gender differences in prevalence
Machine learning

- Knowledge representation and reasoning:
  - how to exploit probabilistic graphical models to capture clinical knowledge
  - model-based diagnosis, prediction and decision-theoretic planning

- Decision support and clinical guidelines: how to integrate task execution with probabilistic reasoning

- Learning probabilistic models about disease interactions from large health-care databases:

  Health-care data
  Nivel, UMC Utrecht
  UMC Nijmegen

  Probabilistic structures
Big healthcare data

- Multiple sources
  - practices

- Source characteristics
  - urbanicity
  - size
  - type

- Statistical methods
  - multi-level
Paired comparison of frequency of occurrence of signs and symptoms given two disorders as likelihood ratio or odds ratio:

\[
\frac{P(f \mid d_1)}{P(f \mid d_2)} \quad \text{or} \quad \frac{\text{Odds}(d_1 \mid f)}{\text{Odds}(d_2 \mid f)}
\]

with \( f \) a feature, e.g. symptom, lab result, and \( d_1, d_2 \) two disorders

Example: Odds Ratios derived from a clinical research:

<table>
<thead>
<tr>
<th></th>
<th>Diabetes Mellitus</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stroke</td>
<td>1.46</td>
</tr>
<tr>
<td>Heart Failure</td>
<td>1.76</td>
</tr>
<tr>
<td>Diabetes Mellitus</td>
<td>1.0</td>
</tr>
<tr>
<td>Hypertension</td>
<td>2.65</td>
</tr>
</tbody>
</table>
Traditional method: regression

- Measures to compare two disorders are determined by:
  - **Linear regression** for continuous outcome $O$ on explanatory variables (predictors) $e$:
    \[ P(O \mid e) \sim \mathcal{N}(\mu, \Sigma) \text{ with } \mu = \mathbb{E}[O \mid e] = \beta^T e \]
  - **Logistic regression** for dichotomous outcomes:
    \[ P(O \mid e) \sim \text{Bernoulli}(p) \text{ with } \logit(\mathbb{E}[O \mid e]) = \beta^T e \]

- Example logistic regression with an interaction term:
  - \[ \logit(\mathbb{E}[F \mid d_1, d_2]) = \beta_0 + \beta_1 d_2 + \beta_2 d_2 + \beta_{12} d_1 d_2 \]
Disorders $D_1$ and $D_2$ and patient findings $F$:

\[
\exp(\beta_{12}) = \frac{\text{Odds}(f \mid d_1, d_2) \text{Odds}(f \mid \overline{d_1}, \overline{d_2})}{\text{Odds}(f \mid \overline{d_1}, \overline{d_2}) \text{Odds}(f \mid d_1, d_2)}
\]

\[
= \frac{\lambda(d_1, d_2 \mid f) \lambda(\overline{d_1}, \overline{d_2} \mid f)}{\lambda(\overline{d_1}, \overline{d_2} \mid f) \lambda(d_1, d_2 \mid f)} \frac{P(d_1, d_2 \mid f) P(\overline{d_1}, \overline{d_2} \mid f)}{P(\overline{d_1}, \overline{d_2} \mid f) P(d_1, d_2 \mid f)}
\]

\[
= \left\{ \frac{P(d_1, d_2 \mid f) P(\overline{d_1}, \overline{d_2} \mid f)}{P(d_1, d_2 \mid f) P(d_1, d_2 \mid f)} \right\} \left\{ \frac{P(\overline{d_1}, \overline{d_2} \mid f) P(d_1, d_2 \mid f)}{P(\overline{d_1}, \overline{d_2} \mid f) P(\overline{d_1}, \overline{d_2} \mid f)} \right\}
\]
Regression gives us outcomes like:

- \( \text{Odds}(d_1 \mid d_2, F) \)
- \( \text{Odds}(d_2 \mid d_1, F) \)

But with some calculation we can also obtain measures like:

- \( \text{Odds}(d_1 \mid F) \)
- \( \text{Odds}(d_2 \mid F) \)
- \( \text{Odds}(d_1, d_2 \mid F) \)

For example, using the odds derived in clinical research on one of the previous slides, we obtain:

\[
\text{Odds}(\text{hypertension, diabetes} \mid \text{heartfailure}) = 1.88
\]
Capturing interaction by regression

Given a set of outcomes and observations, to obtain joint probabilities using regression, in order to investigate interactions within multimorbidity, we need:

- a regression model for each outcome variable of interest
- within each regression model all possible interaction terms
The diagnostic model represents regression analysis of $D_1$. It assumes all remaining variables are independent and certain, whereas in the causal model all true (possible) dependencies are modeled.

Probabilistic graphical models, such as Bayesian networks, support explicit modelling by a graph (uncovered by structure learning).
Concurrent multimorbidity

Independent diseases co-occur at the same time (unconditionally independent)

- \( P(D_i, D_j) = P(D_i)P(D_j) \)

No common signs and symptoms:

- \( \forall F \): Conditional independence
  - \( P(D_i, D_j \mid F) = P(D_i \mid F)P(D_j \mid F) \)
  - Logistic regression: \( \beta_{ij} = 0 \)
  - Structure learning: no edges (paths)

- Depression
- Hypertension
- Symptom
Concurrent multimorbidity

Independent diseases co-occur at the same time (unconditionally independent)

- \( P(D_i, D_j) = P(D_i)P(D_j) \)

Common signs and symptoms:

- \( \exists F \): Conditional dependence
  - \( P(D_i, D_j | F) \neq P(D_i | F)P(D_j | F) \)
  - Logistic regression: \( \beta_{ij} \neq 0 \)
  - Structure learning: \( D_i \rightarrow F \leftarrow D_j \)

Diagram:

- COPD
- Hypertension
- Heart Failure
Dependent diseases:

- \( P(D_i, D_j) \neq P(D_i)P(D_j) \)

because of

- \( \exists F: \text{Common cause} \) (conditional independence)
  - \( P(D_i, D_j \mid F) = P(D_i \mid F)P(D_j \mid F) \)
  - Logistic regression: \( \beta_{ij} = 0 \)
  - Structure learning: \( D_i \leftarrow F \rightarrow D_j \)
Dependent diseases:

- \[ P(D_i, D_j) \neq P(D_i)P(D_j) \]

- \( D_j \) dependent of \( D_i \)
  - \[ P(D_i, D_j \mid F) = P(D_j \mid D_i)P(D_i \mid F) \]
  - Logistic regression: \( \beta_{ij} \neq 0 \)
  - Structure learning: \( F \rightarrow D_i \rightarrow D_j \)

Diagram:

- **Smoking**
- **COPD**
- **Pulmonary Hypertension**
Multimorbidity – types of correlation

Multimorbidity

Concurrent

\[ P(D_i, D_j) = P(D_i)P(D_j) \]

Conditionally Independent

\[ P(D_i, D_j | F) = P(D_i | F)P(D_j | F) \]

\[ \beta_{ij} = 0 \]

Conditionally Dependent

\[ P(D_i, D_j | F) \neq P(D_i | F)P(D_j | F) \]

\[ \beta_{ij} \neq 0 \]

Associative

\[ P(D_i, D_j) \neq P(D_i)P(D_j) \]

Confounding

\[ P(D_i, D_j | F) = P(D_i | F)P(D_j | F) \]

\[ \beta_{ij} = 0 \]

Causal

\[ P(D_i) = P(D_j | D_i)P(D_i) \]

\[ \beta_{ij} \neq 0 \]
Logistic Regression
- Diseases often used as ...
  - outcome variable in one model (A)
  - explanatory variable in another model (B)
  ⇒ multiple models
- Use of interaction terms:
  - $\beta_{ij} = 0 \rightarrow$ True Independence or Confounding?
  - $\beta_{ij} \neq 0 \rightarrow$ Conditional Dependence or Causality?

Bayesian Networks
- All variables treated as uncertain
  - one model!
  - (possible) representation of underlying processes
- Interactions automatically incorporated
- Allows distinguishing between various forms of multimorbidity
Disease modelling by Bayesian networks

**single disease**

- environment
- characteristics
- genetics
- disease
- pathophysiology
- signs
- symptoms
- laboratory results

**multiple diseases**

- environment
- characteristics
- genetics
- disease A
- disease B
- pathophysiology X
- pathophysiology Y
- pathophysiology Z
- sign 1
- sign 2
- sign 3
- symptom 1
- symptom 2
- symptom 3
- laboratory results 1
- laboratory results 2
- laboratory results 3

Abstract model of a single disease (left) and multiple diseases (right)
To model variation of outcomes between various groups (e.g. different general practices), taking into account correlation within groups

Formulation in terms of regression models (with \( l \) being a vector of higher level variables):

- multilevel linear regression:
  
  \[
  P(O_k | e, l) \sim \mathcal{N}(\mu, \Sigma) \text{ with } \mu = \mathbb{E}[O | e, l] = \beta_k e = (\delta_k + \gamma_k l)^T e
  \]

- multilevel logistic regression:
  
  \[
  P(O_k | e, l) \sim \text{Bernoulli}(p) \text{ with } \text{logit} (\mathbb{E}[O_k | e, l]) = (\delta_k + \gamma_k l)^T e
  \]

with \( k \) the group index, \( \gamma_k \) the level parameters, and \( \delta_k \) the group variation
Limitations:

- Only comparison between one outcome variable and predictors
- Only predictions are treated as uncertain
- No explicit knowledge about relationships between predictors
- Within multimorbidity some variables are both outcome and predictor
Disease modelling of multimorbidity

Graphical representation of risks, pathophysiology, and symptomatology:

- Environment
- Patient characteristics
- Genetics
- Therapy $T_A$
- Therapy $T_C$
- Physiology $P_X$
- Physiology $P_Y$
- Symptomatology $S_X$
- Symptomatology $S_Y$
- Geriatric Syndrome
- Daily functioning

Disease

Pathophysiology $A$

Pathophysiology $B$

Pathophysiology $C$
Disease modelling of multimorbidity

Graphical representation of risks, pathophysiology, and symptomatology:

- $\text{logit}(E[DiseaseA \mid \text{Age, Gender, SymptomX}]) = \beta_0A + \beta_{1A}\text{Age} + \beta_{2A}\text{Gender} + \beta_{3A}\text{SymptomX}$
- $\text{logit}(E[DiseaseB \mid \text{Age, Gender, SymptomX, SymptomY}]) = \beta_0B + \beta_{1B}\text{Age} + \beta_{2B}\text{Gender} + \beta_{3B}\text{SymptomX} + \beta_{4B}\text{SymptomY}$
- $\text{logit}(E[DiseaseC \mid \text{Age, Gender, SymptomY, DiseaseB}]) = \beta_0C + \beta_{1C}\text{Age} + \beta_{2C}\text{Gender} + \beta_{3C}\text{SymptomX} + \beta_{4C}\text{DiseaseB}$
A **Bayesian network** is a tuple $\mathcal{B} = (G, X, P)$, with $G = (V, E)$ a directed acyclic graph, $X = \{X_v \mid v \in V\}$ a set of random variables indexed by $V$, and $P$ a joint probability distribution such that:

$$P(X_1 = x_1 \land \cdots \land X_n = x_n) = \prod_{v \in V} P(X_v = x_v \mid X_j = x_j \text{ for all } j \in \pi(v))$$

**Simple example:**

\[
\begin{align*}
G & \quad \downarrow \\
HT & \quad DM
\end{align*}
\]

$←$ explanatory variables

$←$ outcome variables

$$P(V) = P(X_{HF} \mid X_{HT}, X_{DM})P(X_{HT} \mid X_G)P(X_{DM} \mid X_G)P(X_G)$$

**Structure and parameters** of a Bayesian network can be learned from data.
Combination of concepts

In summary, with patient data acquired from general practices and the aim of modelling multiple disease, we are facing:

1. hierarchical data structures
   → which can be analysed using multilevel regression

2. multiple diseases with multiple possible interactions
   → which can be modelled using probabilistic graphical methods
   - Bayesian networks
   - undirected graphs
   - hybrid graphs

Our goal → adopting both concepts into multilevel Bayesian networks
MLBN with independence and intra-level structure

- Here all variables are uncertain (random) and expressed as such.
- Representation of different levels of outcomes (and other variables).
- Inter-level dependence $\rightarrow$
- Intra-level dependence $\rightarrow$

![Diagram of MLBN with independence and intra-level structure]
Transition probabilities

# health determinants
- **Base line**
  - 0: 86%
  - 1: 81%
  - 2: 86%
  - 3: 94%
  - 4+: 100%

- **3 years**
  - 0: 12%
  - 1: 17%
  - 2: 2%
  - 3: 3%
  - 4+: 100%

- **5 years**
  - 0: 8%
  - 1: 11%
  - 2: 2%
  - 3: 3%
  - 4+: 100%

# chronic cardiovascular diseases
- **Base line**
  - 0: 87%
  - 1: 73%
  - 2: 73%
  - 3: 68%
  - 4+: 60%

- **3 years**
  - 0: 11%
  - 1: 22%
  - 2: 25%
  - 3: 25%
  - 4+: 55%

- **5 years**
  - 0: 11%
  - 1: 22%
  - 2: 25%
  - 3: 25%
  - 4+: 55%
In context - diabetes mellitus

- Ischemic heart disease
- Heart failure
- Stroke
- Nephropathy

Diabetics
- Base line
- 3yr follow-up
- 5yr follow-up

Non-diabetics
- Base line
- 3yr follow-up
- 5yr follow-up
Conclusions

Machine learning in medicine
- Requires a combination of knowledge representation, reasoning and learning methods
- Big healthcare data: need for new methods

Methodology
- Integration of multilevel analysis and Bayesian networks
- Visualization of interactions between disease variables
- Personalization of patients (e.g., diabetics)
- Fundament towards clinical guidelines that deal with multimorbidity