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
- $\Rightarrow$ Machine learning should take knowledge representation and uncertainty into account (compare medical statistics)

Pandora's box

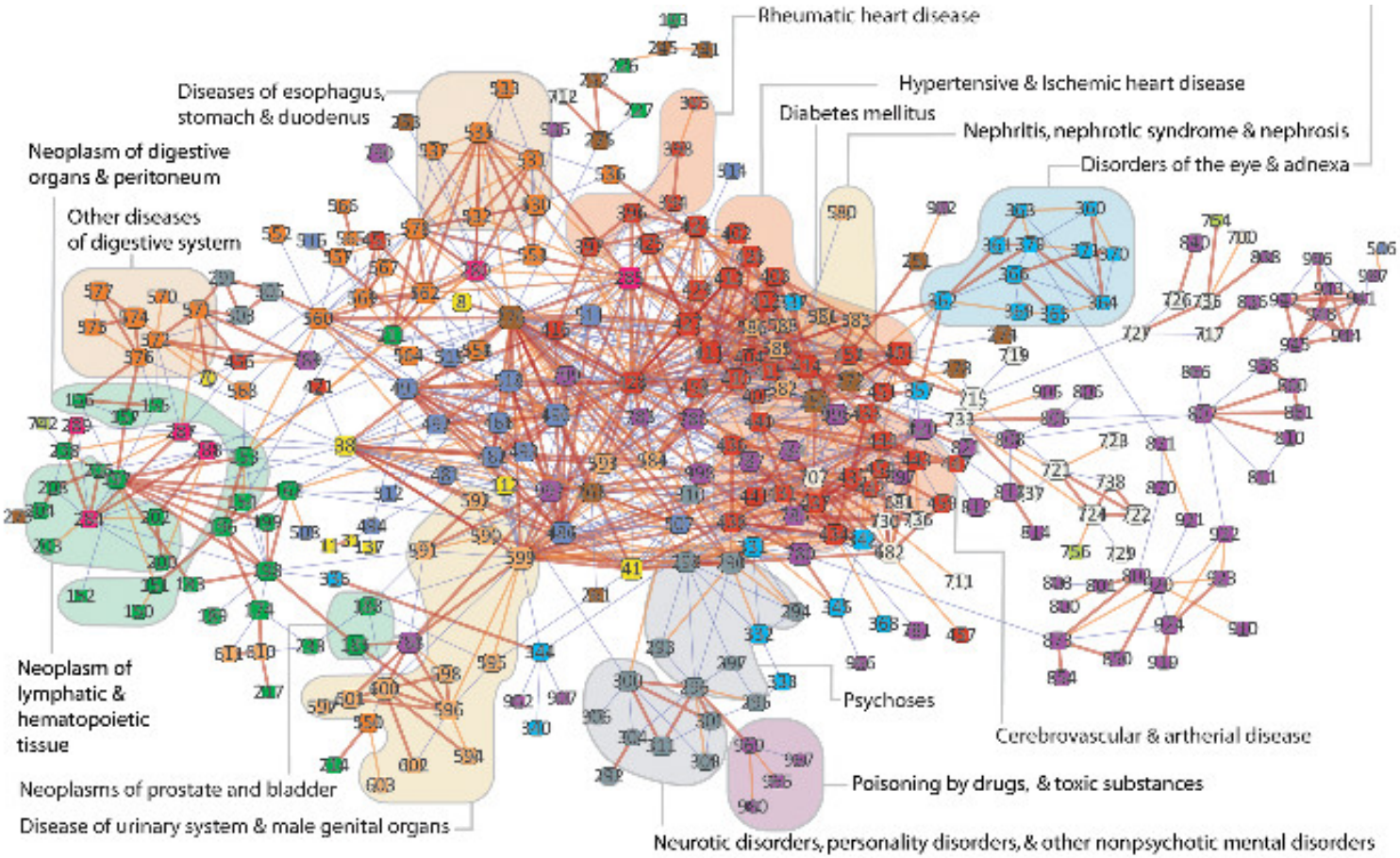
Rob is 67 years old and has been ill for some time, in particular he is currently treated for:

- 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

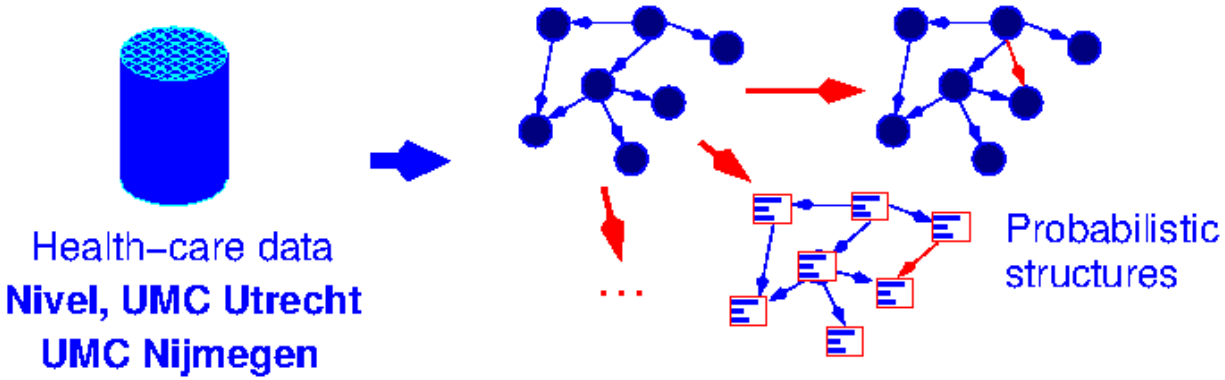
Challenge: dealing with diseases and their interactions



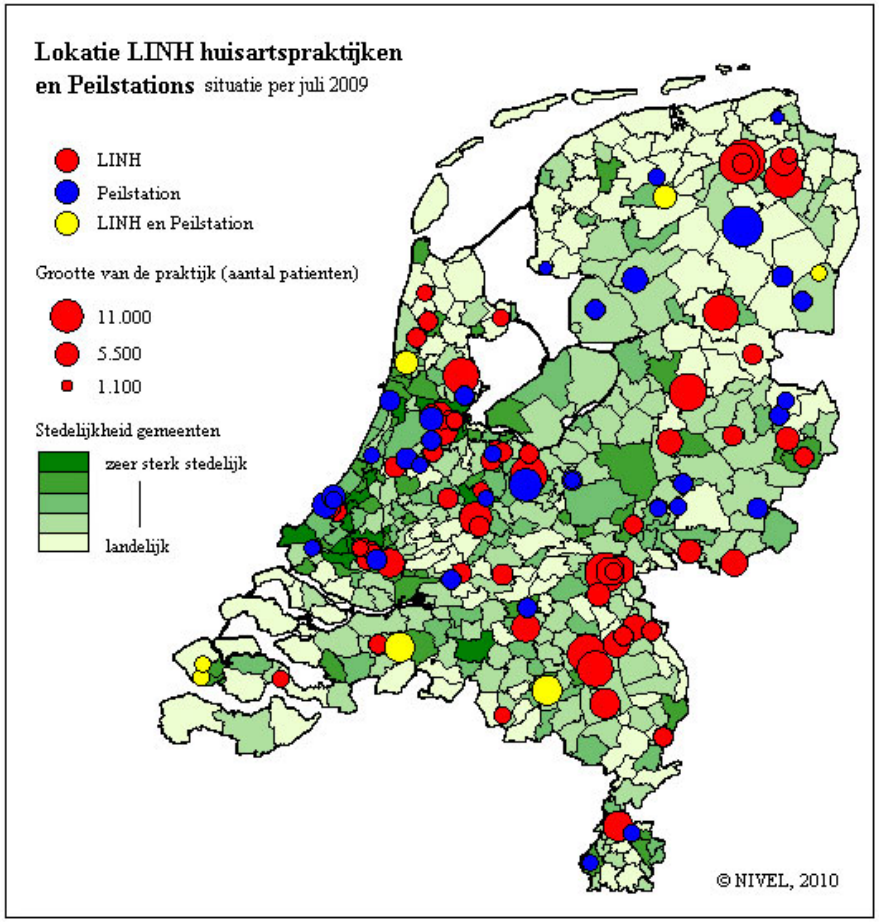
- Complexity: many individual diseases and classes of disease
- Probabilistic relationships: uncertain interactions between diseases, regional, social, and gender differences in prevalence

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:



Big healthcare data



- Multiple sources
  - practices
- Source characteristics
  - urbanicity
  - size
  - type
- Statistical methods
  - multi-level

State of the art in multimorbidity

- Paired comparison of frequence of occurence of signs and symptoms given two disorders as likelihood ratio or odds ratio:

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

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

- Example: Odds Ratios derived from a clinical research:

	Diabetes Mellitus
Stroke	1.46
Heart Failure	1.76
Diabetes Mellitus	1.0
Hypertension	2.65

Traditional method: regression

- Measures to compare two disorders are determined by:
  - Linear regression for continuous outcome O on explanatory variables (predictors) e:

    $P(O \mid e) \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \mathrm{E}[O \mid e] = \beta^T e$

  - Logistic regression for dichotomous outcomes:

    $P(O \mid e) \sim \text{Bernoulli}(p)$ with $\text{logit}(\mathrm{E}[O \mid e]) = \beta^T e$

- Example logistic regression with an interaction term:
  - $\text{logit}(\mathrm{E}[F \mid d_1, d_2]) = \beta_0 + \beta_1 d_2 + \beta_2 d_2 + \beta_{12} d_1 d_2$

Regression: interaction term β_{ij}

Disorders D_1 and D_2 and patient findings F:

$$
\begin{aligned}
exp(\beta_{12}) &= \frac{Odds(f \mid d_1, d_2)Odds(f \mid \overline{d_1}, \overline{d_2})}{Odds(f \mid \overline{d_1}, d_2)Odds(f \mid d_1, \overline{d_2})} \\
&= \frac{\lambda(d_1, d_2 \mid f)\lambda(\overline{d_1}, \overline{d_2} \mid f)}{\lambda(\overline{d_1}, d_2 \mid f)\lambda(d_1, \overline{d_2} \mid f)} \\
&= \frac{\frac{P(d_1,d_2|f)}{P(d_1,d_2|\overline{f})}\frac{P(\overline{d_1},\overline{d_2}|f)}{P(\overline{d_1},\overline{d_2}|\overline{f})}}{\frac{P(\overline{d_1},d_2|f)}{P(\overline{d_1},d_2|\overline{f})}\frac{P(d_1,\overline{d_2}|f)}{P(d_1,\overline{d_2}|\overline{f})}} \\
&= \left\{\frac{P(d_1, d_2 \mid f)P(\overline{d_1}, \overline{d_2} \mid f)}{P(d_1, \overline{d_2} \mid f)P(\overline{d_1}, d_2 \mid f)}\right\}\left\{\frac{P(\overline{d_1}, d_2 \mid \overline{f})P(d_1, \overline{d_2} \mid \overline{f})}{P(d_1, d_2 \mid \overline{f})P(\overline{d_1}, \overline{d_2} \mid \overline{f})}\right\}
\end{aligned}
$$

Regression: other measures

Regression gives us outcomes like:

- $Odds(d_1 \mid d_2, F)$
- $Odds(d_2 \mid d_1, F)$

But with some calculation we can also obtain measures like:

- $Odds(d_1 \mid F)$
- $Odds(d_2 \mid F)$
- $Odds(d_1, d_2 \mid F)$

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

$$Odds(hypertension, diabetes \mid heartfailure) = 1.88$$

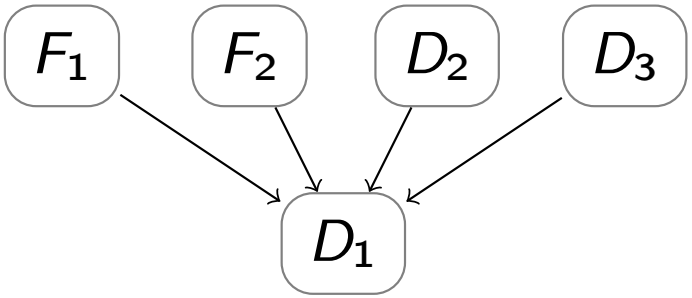
Capturing interaction by regression

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

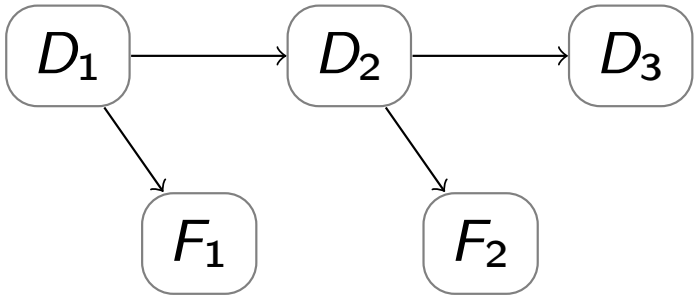
- a regression model for each outcome variable of interest
- within each regression model all possible interaction terms

Graphical representation

wrong, diagnostic model



right, causal model



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)

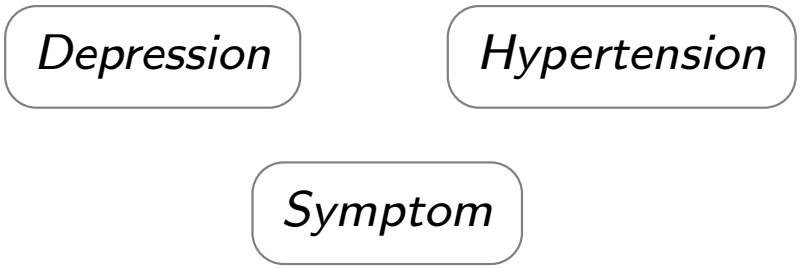
Concurrent multimorbidity

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)



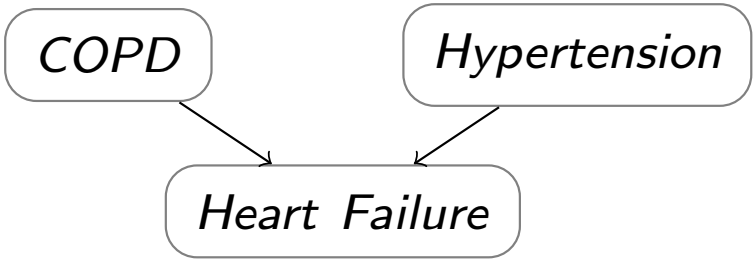
Concurrent multimorbidity

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
  - $P(D_i, D_j \mid F) \neq P(D_i \mid F)P(D_j \mid F)$
  - Logistic regression: $\beta_{ij} \neq 0$
  - Structure learning: $D_i \rightarrow F \leftarrow D_j$



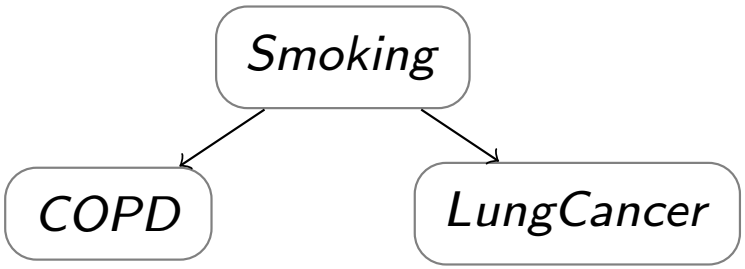
Confounding – common cause – multimorbidity

Dependent diseases:

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

because of

- $\exists F$: 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: $D_i \leftarrow F \rightarrow D_j$

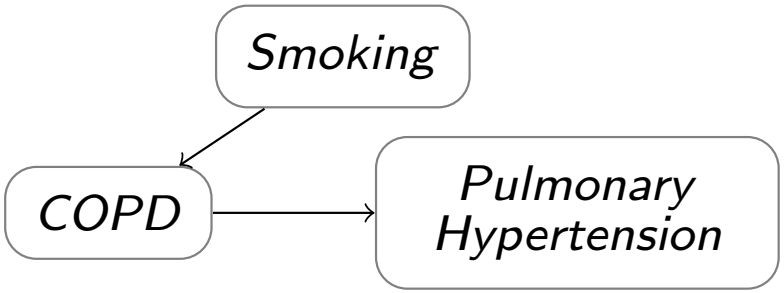


Causal multimorbidity

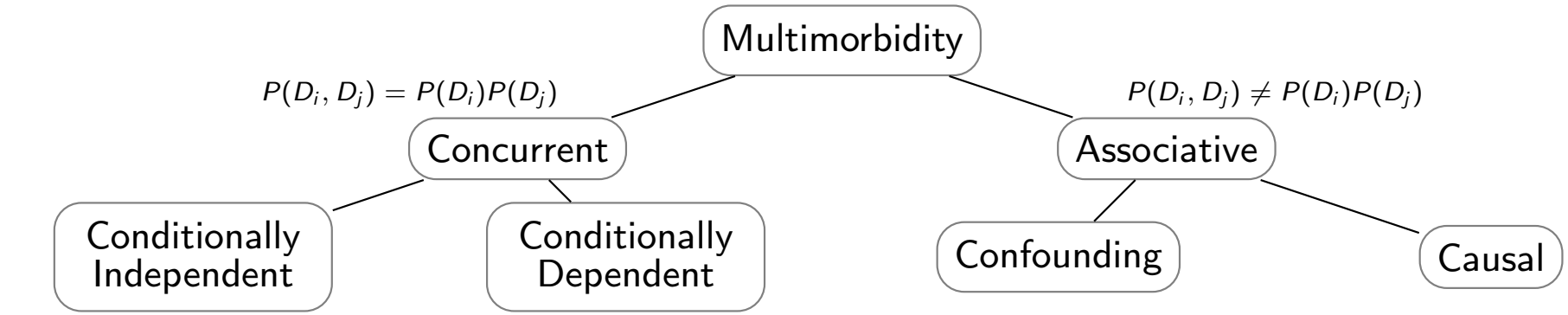
Dependent diseases:

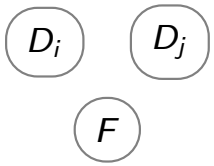
- $P(D_i, D_j) \neq P(D_i)P(D_j)$

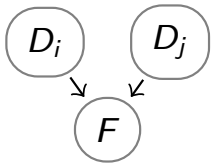
- D_j dependent of D_i
  - $P(D_i, D_j \mid F) = P(D_j \mid D_i)P(D_i \mid F)$
  - Logistic regression: $\beta_{ij} \neq 0$
  - Structure learning: $F \rightarrow D_i \rightarrow D_j$

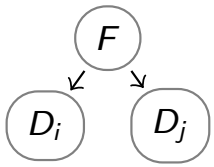


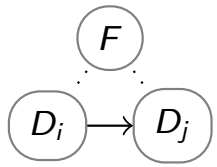
Multimorbidity – types of correlation









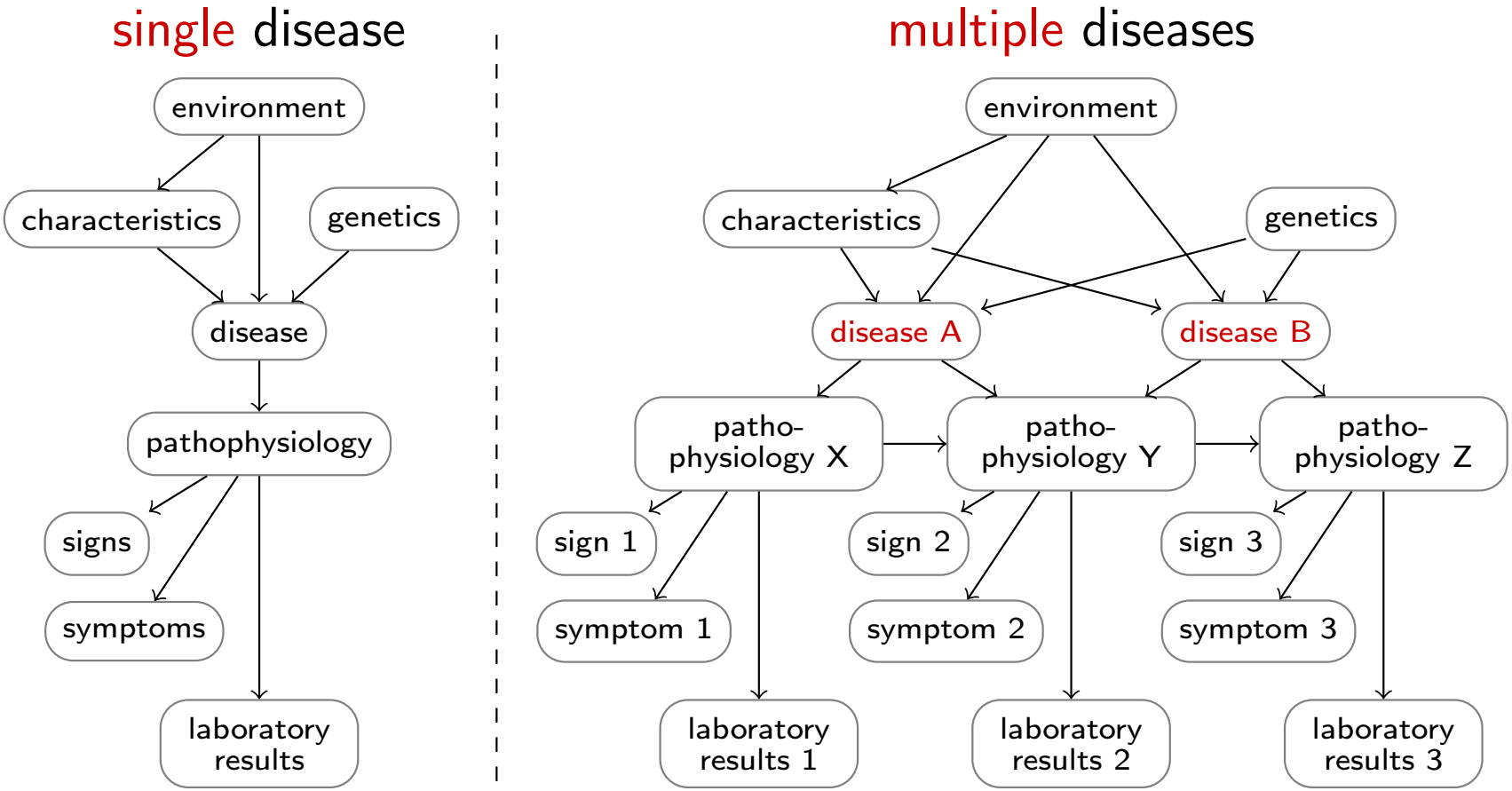


$\beta_{ij} = 0$ $\beta_{ij} \neq 0$ $\beta_{ij} = 0$ $\beta_{ij} \neq 0$

Logistic regression versus Bayesian networks

- Logistic Regression
  - Diseases often used as ...
    - outcome variable in one model (A)
    - explanatory variable in another model (B)
    - $\Rightarrow$ 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



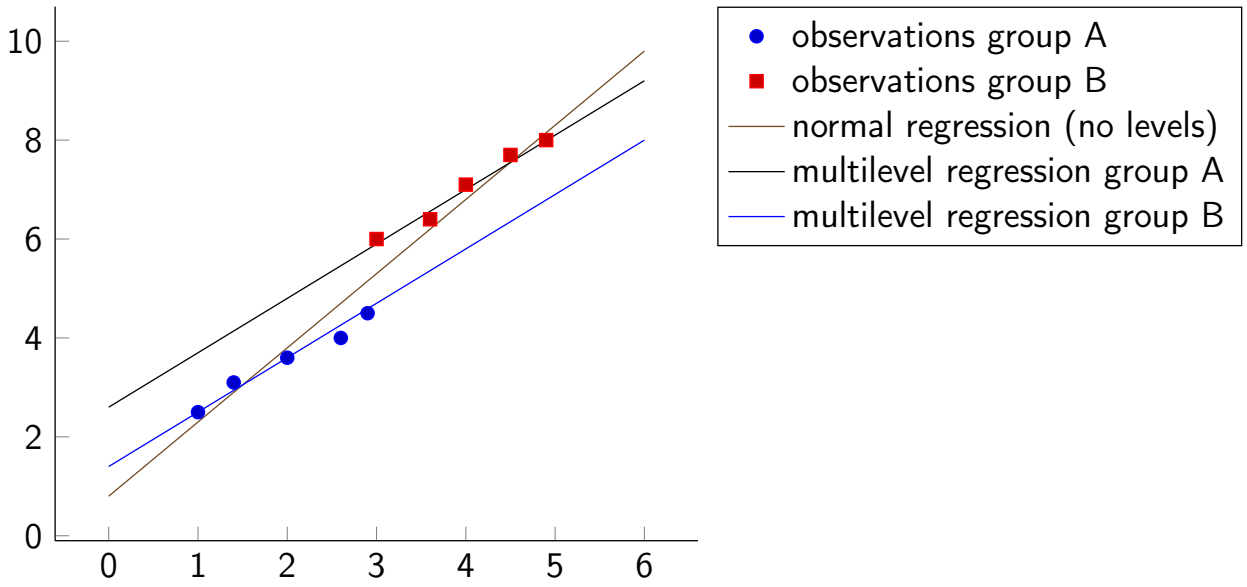
Abstract model of a single disease (left) and multiple diseases (right)

Big data: multilevel regression

- 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 I being a vector of higher level variables):
  - multilevel linear regression:
    $P(O_k \mid e, I) \sim \mathcal{N}(\mu, \Sigma)$ with $\mu = \mathrm{E}[O \mid e, I] = \beta_k e = (\delta_k + \gamma_k I)^T e$
  - multilevel logistic regression:
    $P(O_k \mid e, I) \sim \mathrm{Bernoulli}(p)$ with $\mathrm{logit}(\mathrm{E}[O_k \mid e, I]) = (\delta_k + \gamma_k I)^T e$

  with k the group index, γ_k the level parameters, and δ_k the group variation

Multilevel regression

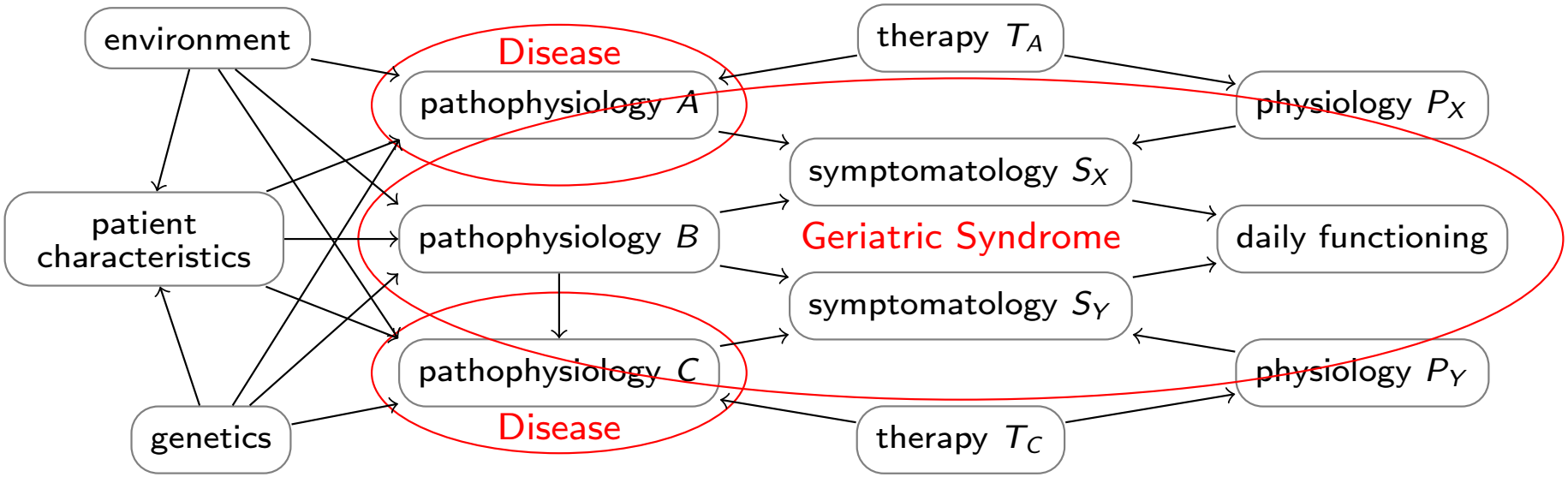


Limitations:

- 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

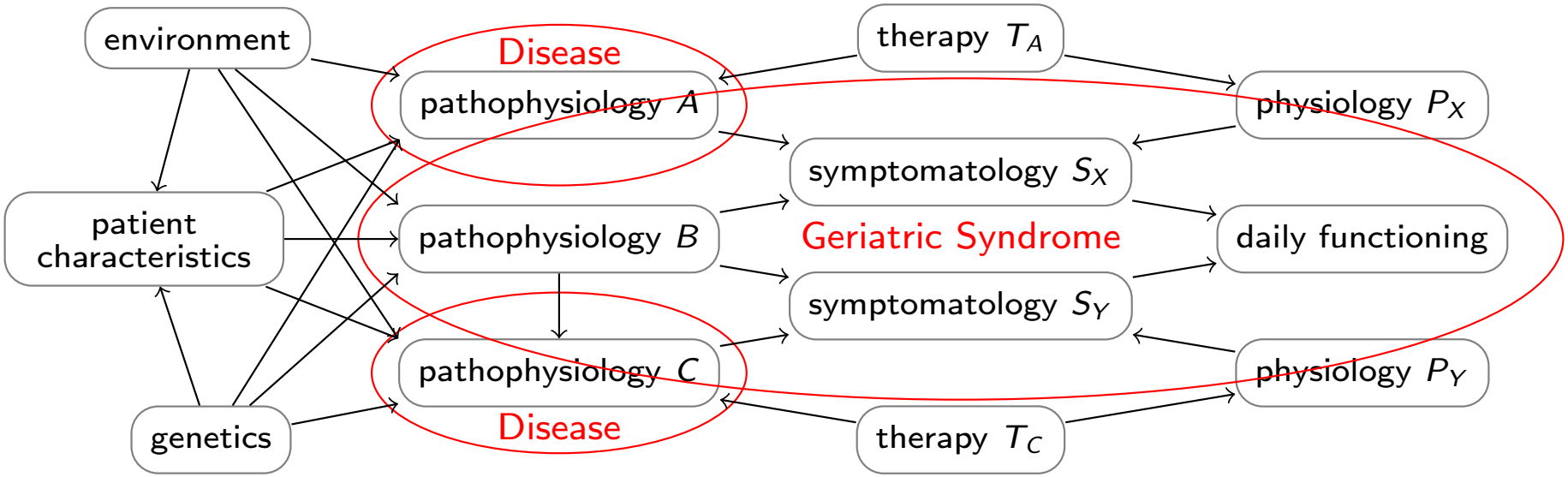
Disease modelling of multimorbidity

Graphical representation of risks, pathophysiology, and symptomatology:



Disease modelling of multimorbidity

Graphical representation of risks, pathophysiology, and symptomatology:



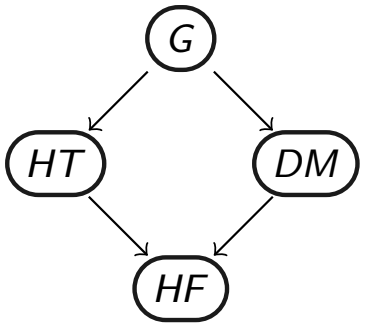
- $\text{logit}(\text{E}[\mathit{DiseaseA} \mid \mathit{Age}, \mathit{Gender}, \mathit{SymptomX}]) = \beta_{0A} + \beta_{1A}\mathit{Age} + \beta_{2A}\mathit{Gender} + \beta_{3A}\mathit{SymptomX}$
- $\text{logit}(\text{E}[\mathit{DiseaseB} \mid \mathit{Age}, \mathit{Gender}, \mathit{SymptomX}, \mathit{SymptomY}]) = \beta_{0B} + \beta_{1B}\mathit{Age} + \beta_{2B}\mathit{Gender} + \beta_{3B}\mathit{SymptomX} + \beta_{4B}\mathit{SymptomY}$
- $\text{logit}(\text{E}[\mathit{DiseaseC} \mid \mathit{Age}, \mathit{Gender}, \mathit{SymptomY}, \mathit{DiseaseB}]) = \beta_{0C} + \beta_{1C}\mathit{Age} + \beta_{2C}\mathit{Gender} + \beta_{3C}\mathit{SymptomX} + \beta_{4C}\mathit{DiseaseB}$

As a Bayesian network

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:

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

Simple example:



← 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:

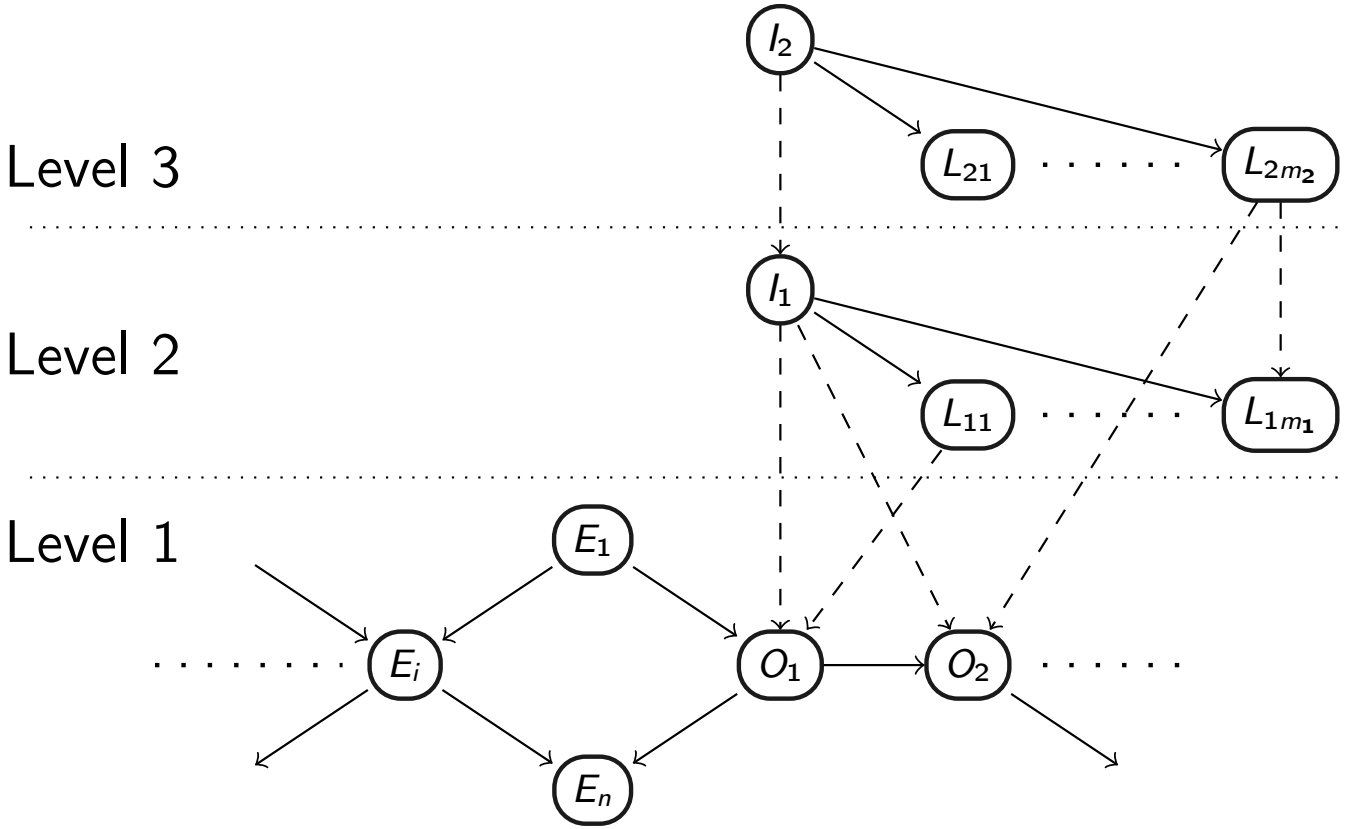
1. hierarchical data structures
   $\rightarrow$ which can be analysed using multilevel regression

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

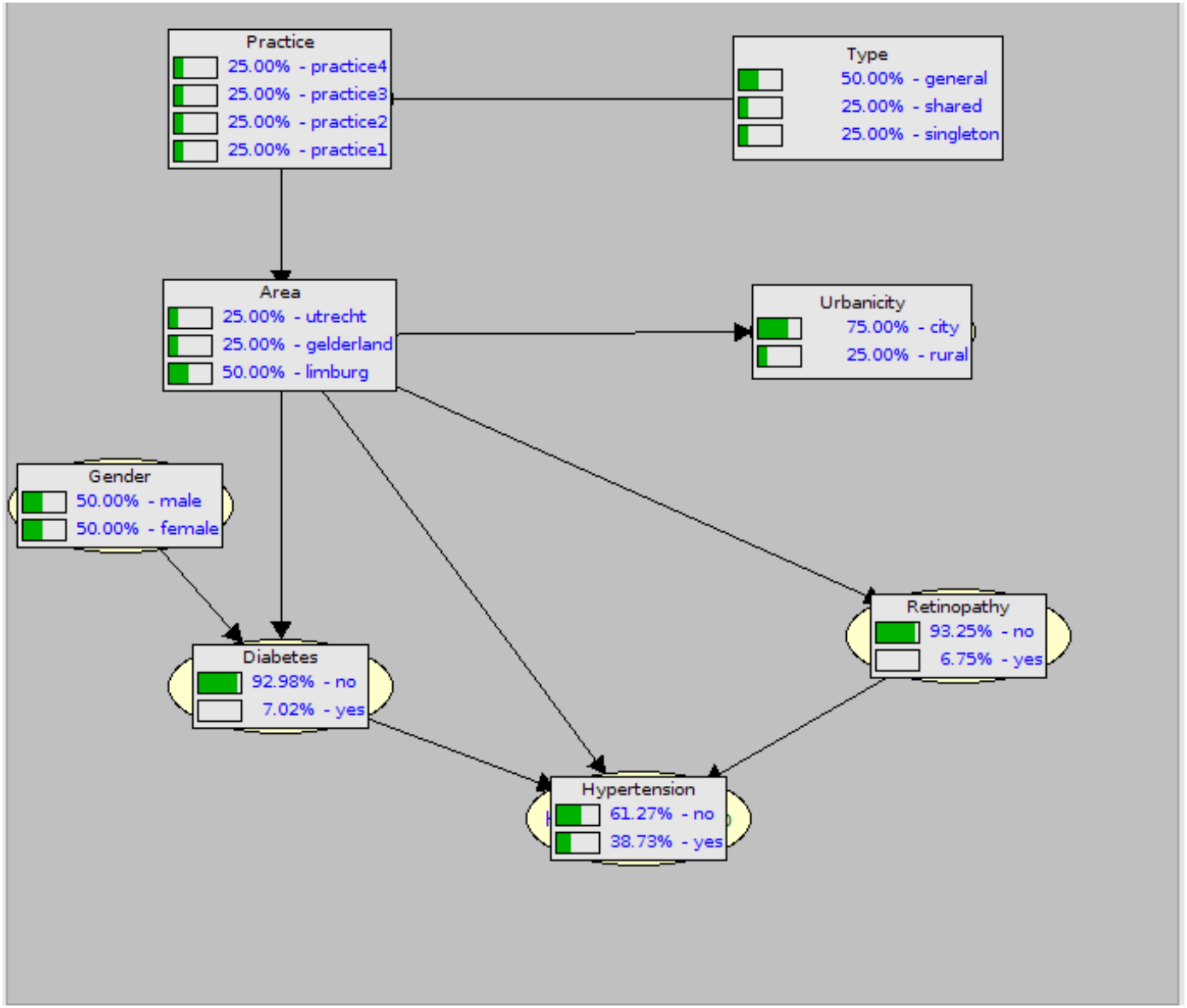
Our goal $\rightarrow$ adopting both concepts into multilevel Bayesian networks

MLBN with independence and intra-level structure

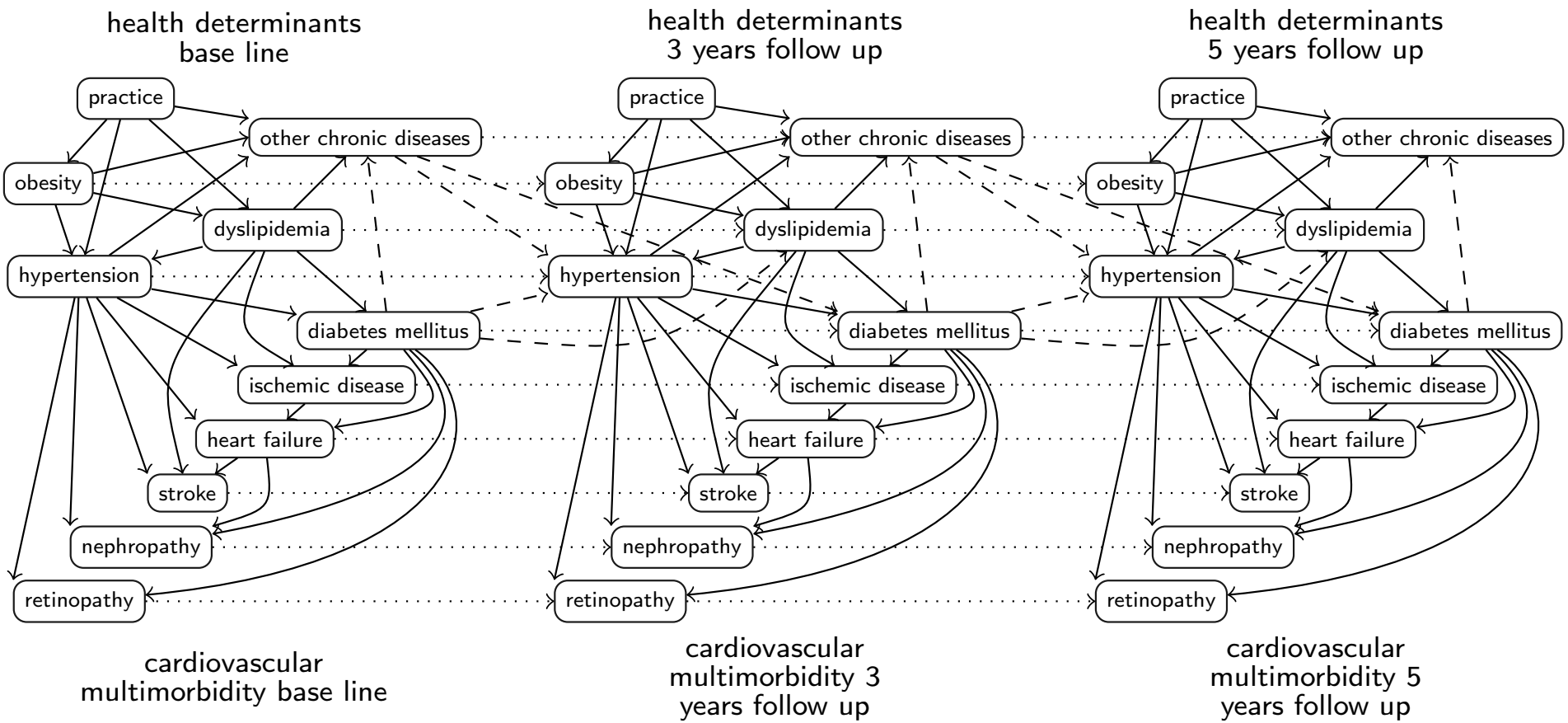
- Here all variables are uncertain (random) and expressed as such
- Representation of different levels of outcomes (and other variables)
- Inter-level dependence $\dashrightarrow$
- Intra-level dependence $\rightarrow$



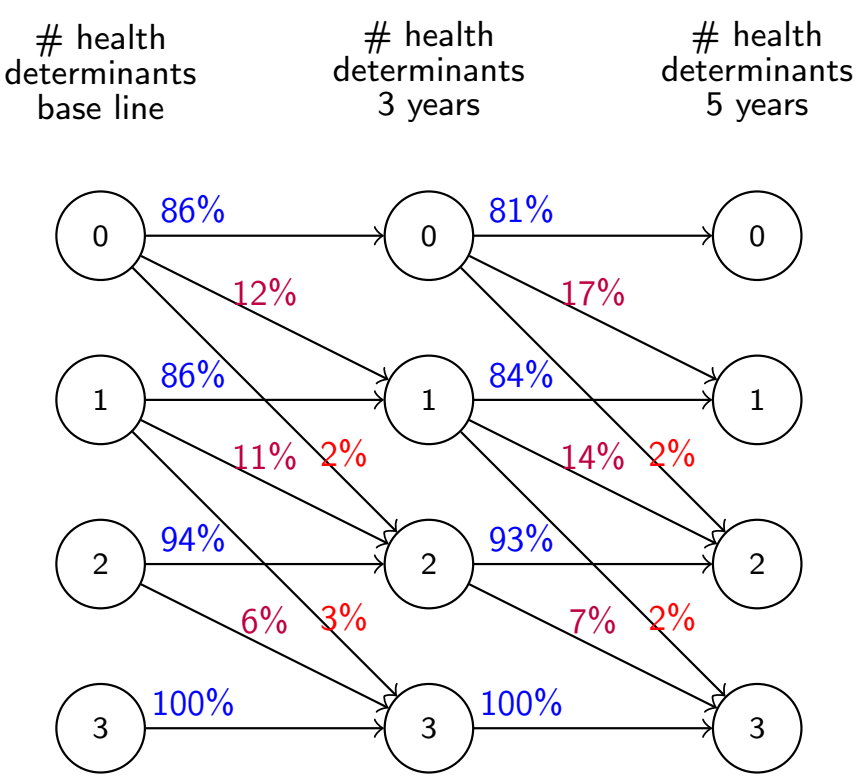
Toy example

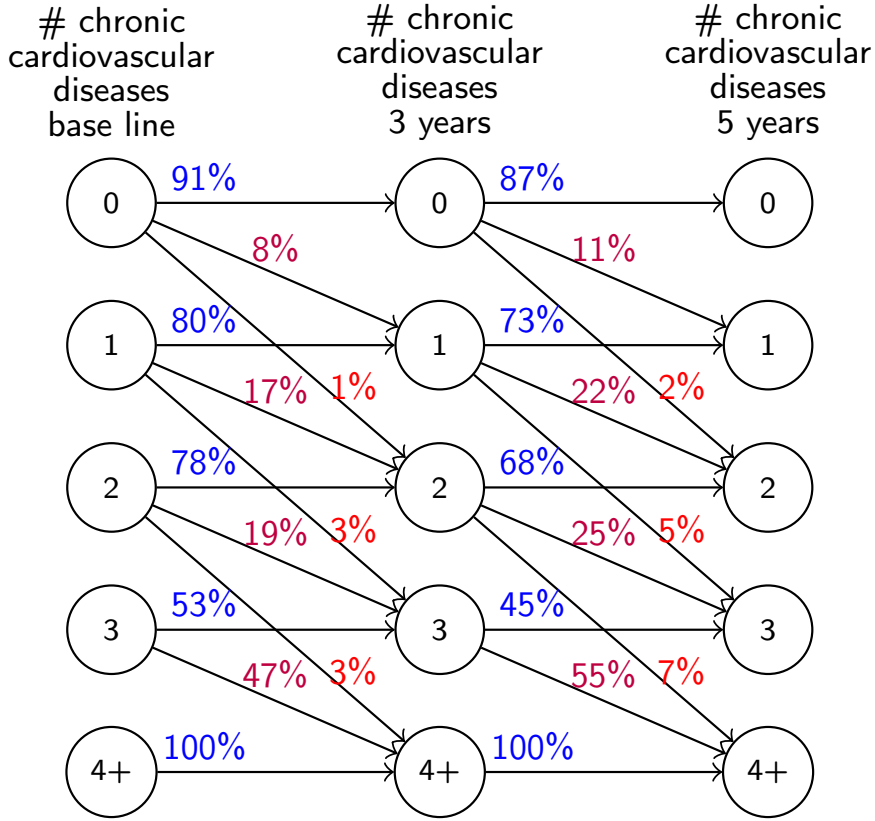

Practice
25.00% - practice4
25.00% - practice3
25.00% - practice2
25.00% - practice1
Type
50.00% - general
25.00% - shared
25.00% - singleton
Area
25.00% - utrecht
25.00% - gelderland
50.00% - limburg
Urbanicity
75.00% - city
25.00% - rural
Gender
50.00% - male
50.00% - female
Retinopathy
93.25% - no
6.75% - yes
Diabetes
92.98% - no
7.02% - yes
Hypertension
61.27% - no
38.73% - yes

Cardiovascular model - MLBN at 3 time points

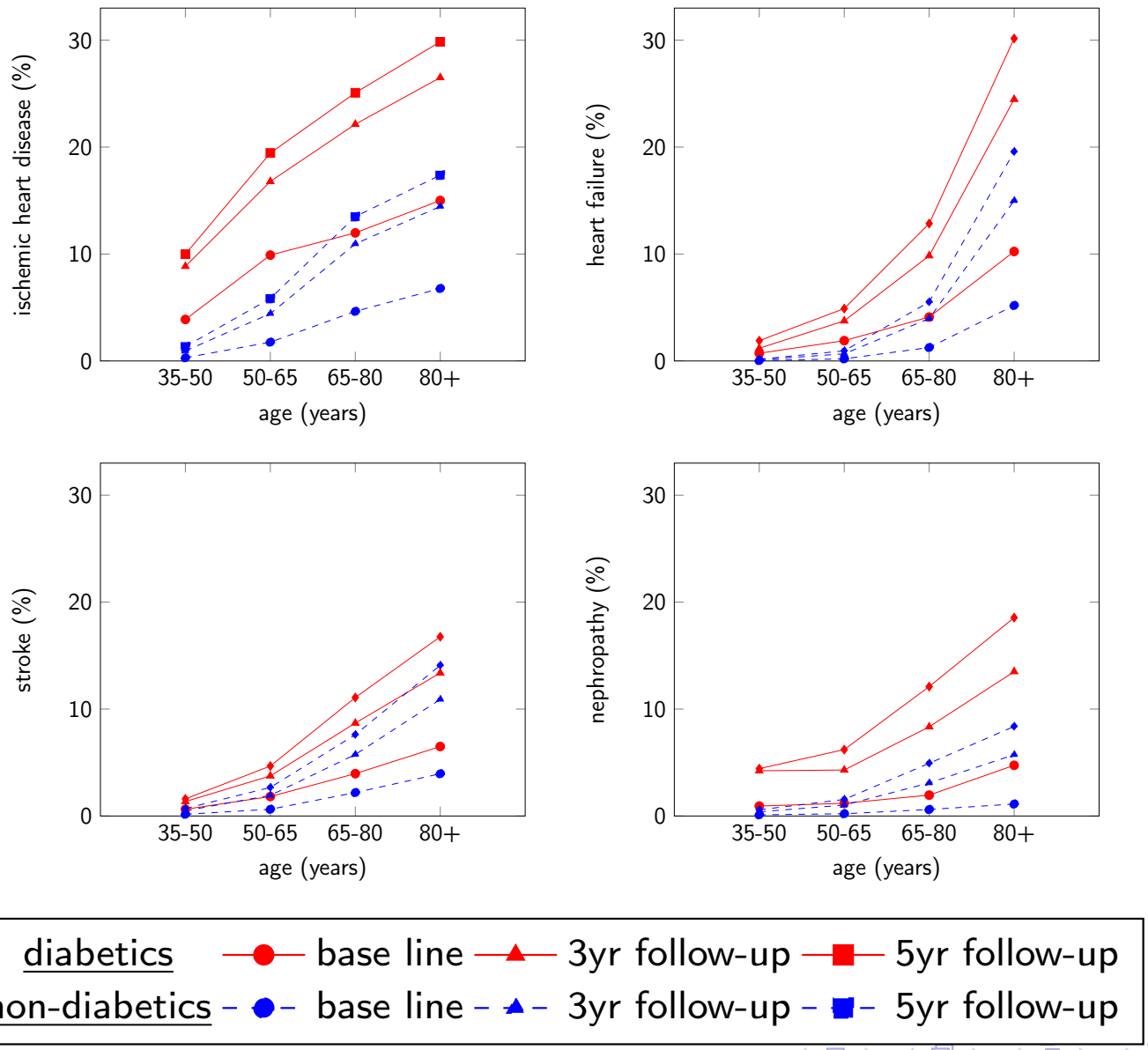


Transition probabilities





In context - diabetes mellitus



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