Towards building a Bayesian Network COPD model

PMI RESEARCH & DEVELOPMENT

6th International Conference on Computational Management Science; May 1-3, 2009; Geneva, Switzerland
Peter Sperisen

Outline

- Context
  - Bayesian Networks and Disease Biology
  - Objectives
- Data
  - Sources and Processing
- Case Study (Human *in vivo* sub-model)
  - From Cigarette Smoke to FEV1
- Hugin Model
- Issues
- Next Steps

Context - Bayesian Networks (BN)

- A Bayesian network (or a belief network) is a probabilistic graphical model (directed acyclic graph) that represents a set of variables and their probabilistic independencies.

- Nodes can represent any kind of variable, be it a measured parameter, a latent variable, or a hypothesis. They are not restricted to representing random variables. Links, on the other hand, represent causal relationships.

- The network structure of a BN can either be learned from data (bottom-up) or specified by experts (top-down).

- In order to fully specify the Bayesian network and thus fully represent the joint probability distribution, it is necessary to specify for each node X the probability distribution for X conditional upon X's parents.

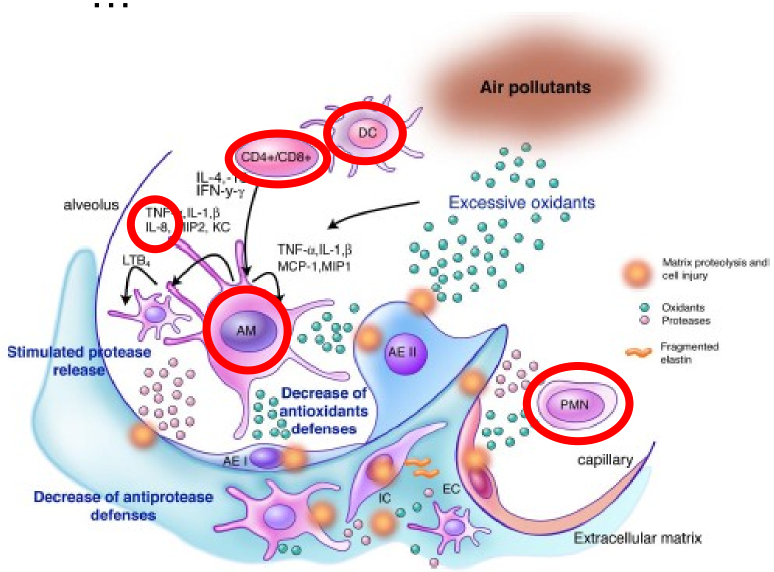
(adapted from Wikipedia)

Context - Disease Biology

- Chronic obstructive pulmonary disease (COPD) is a chronic respiratory disease characterized by airflow limitation that is poorly reversible with bronchodilators (unlike asthma). The airflow limitation usually gets progressively worse over time.
- COPD is strongly linked to exposure to noxious particles or gases, such as cigarette smoke, which trigger an abnormal inflammatory response in the lung.
- The inflammatory response in the larger airways is known as **chronic bronchitis**, whereas in the alveoli, the inflammatory response causes the progressive destruction of the lung tissue known as **emphysema**.
- The natural course of COPD is characterized by the occasional sudden worsening of symptoms called acute exacerbations, caused mainly by infections or air pollution.

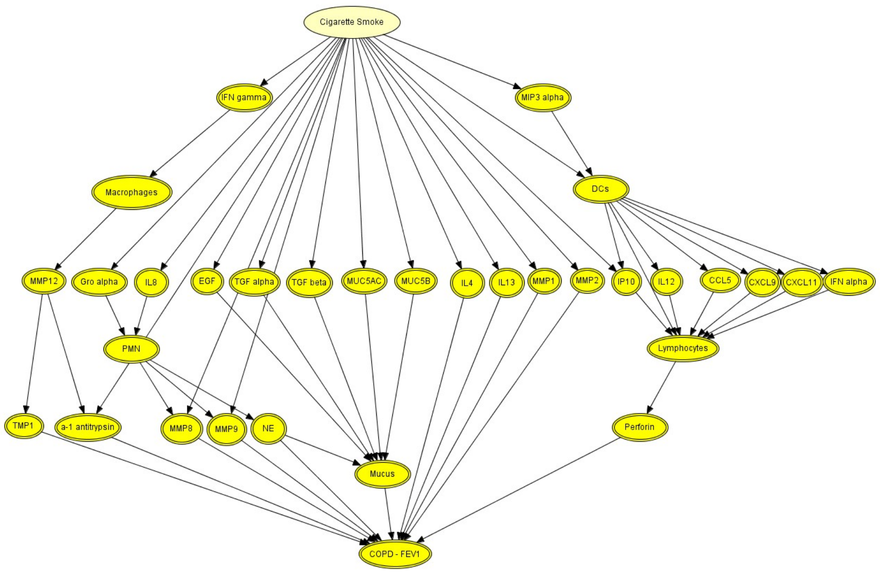
Context – From Disease to a BN COPD Model

- In-house COPD experts provided us with the main pathways involved in the development of COPD:
  - CS → Epithelium → IL-8 → PMNs → NE, MMP8, MMP9 → COPD/FEV1
  - CS → Macrophages → IL-8 → PMNs → NE, MMP8, MMP9 → COPD/FEV1
  - …



Yoshida & Tuder, 2007 (Fig. 2)

Main pathways



Context – Objectives

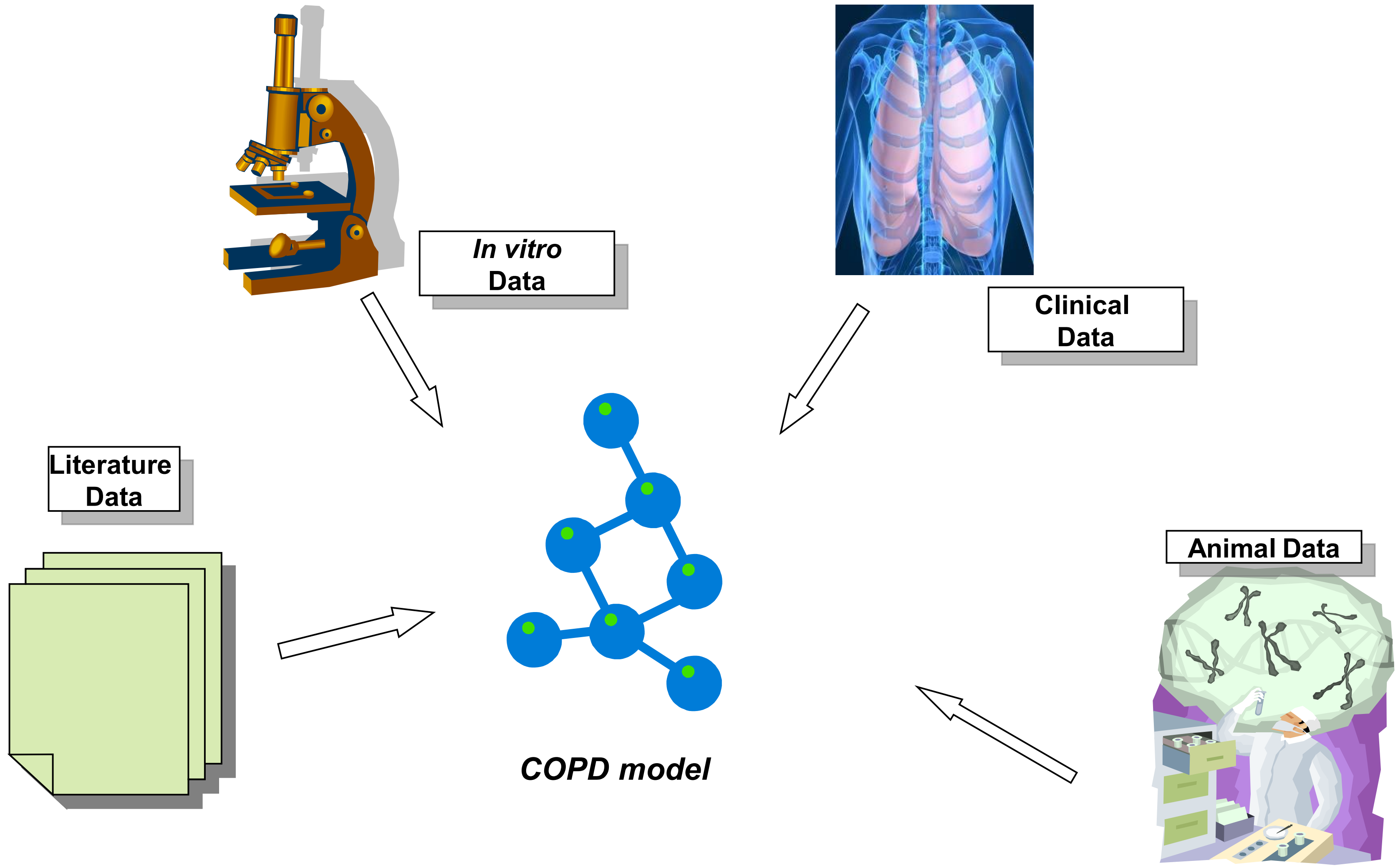
- Objectives of COPD modeling:

  - … predict the risk of COPD associated with cigarette smoke in the absence epidemiological studies

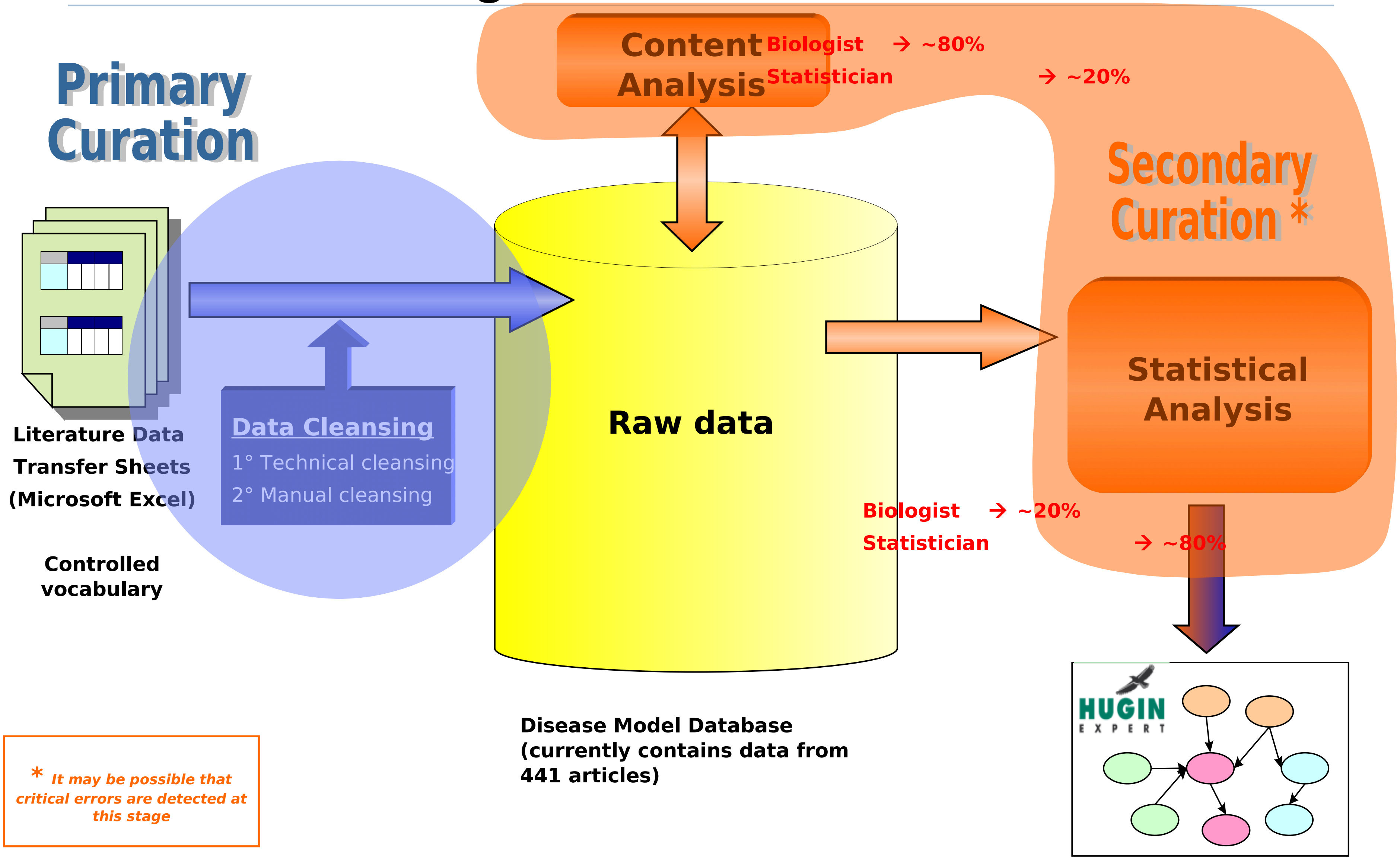
  - … better understand the disease mechanisms of COPD

  - … identify the most relevant biomarkers of COPD

Data - Sources



Data - Processing



Case Study - Strategy of Analysis

Case 1

Link categorical variable → continuous variable

Case 2

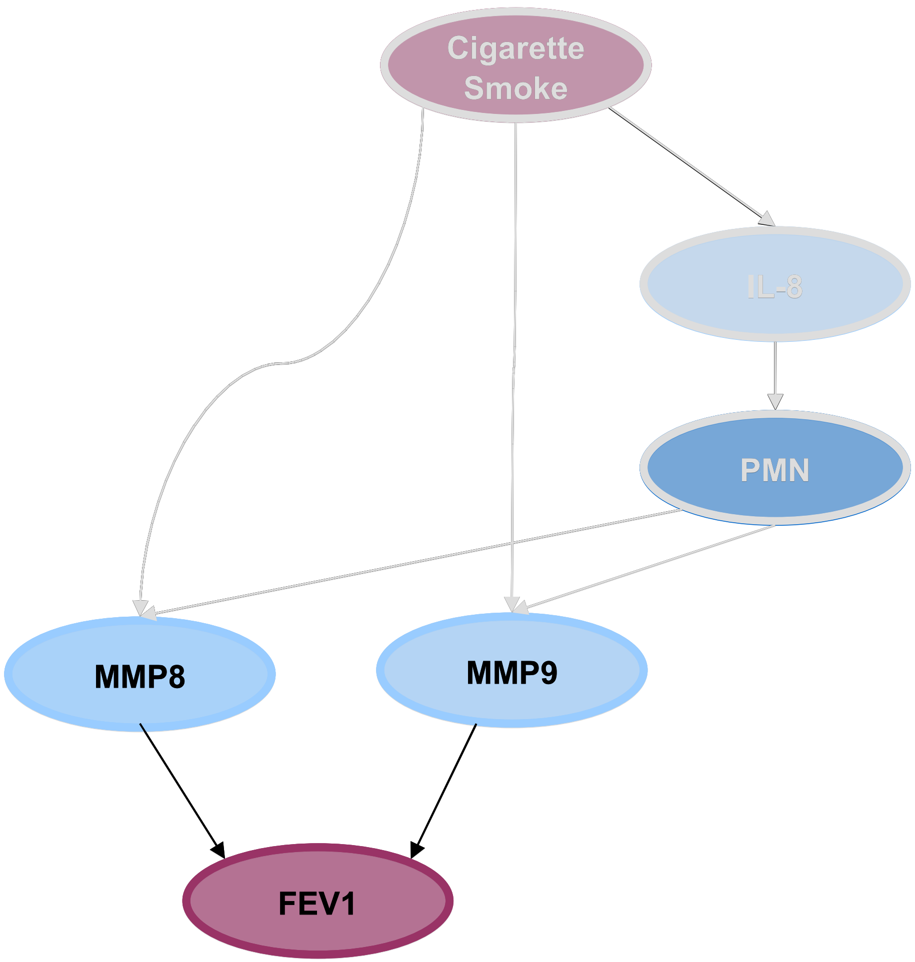
Link continuous variable → continuous variable

Case 3

Link categorical + continuous variables → continuous variable

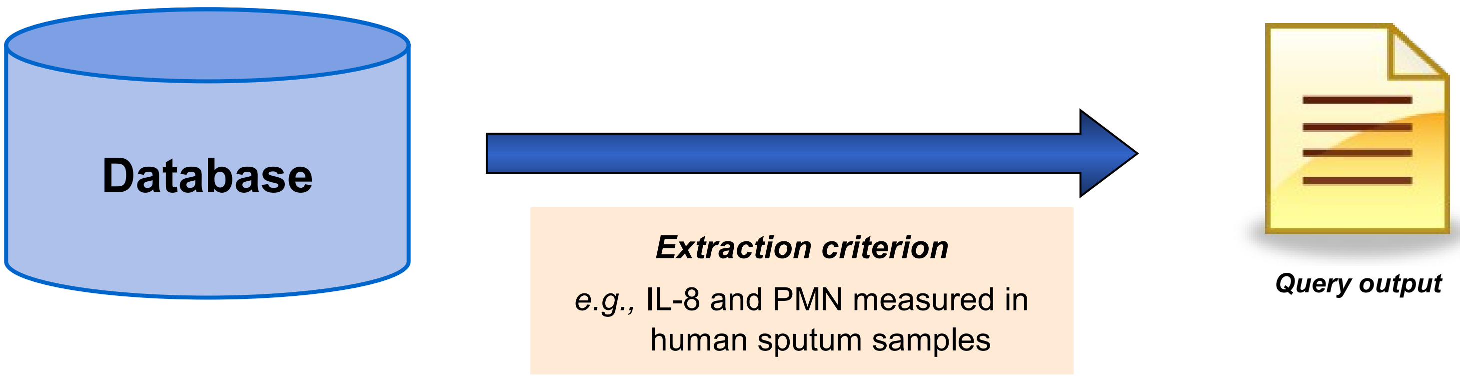
Case 4

Link (at least) 2 continuous variables → continuous variable



Case Study - Preliminaries

- Description of the data extraction process…



- Diversity in extracted data…
  - Different smoking status: *smoker, non-smoker, ex-smoker, smoker + no smoker, …*
  - Different groups: *healthy, COPD, chronic bronchitis, asthma, …*
  - Different units: *ng/mL, nMolar, nMol, %, pg/mL / mg albumin, …*
  - Different statistics: *mean, std, median, inter-quartile range, …*
  - Correlations and Regressions
  - Scatterplots of individual values

Case Study – Inclusion Criteria

- Human model
  - Whenever possible, use only human *in vivo* data

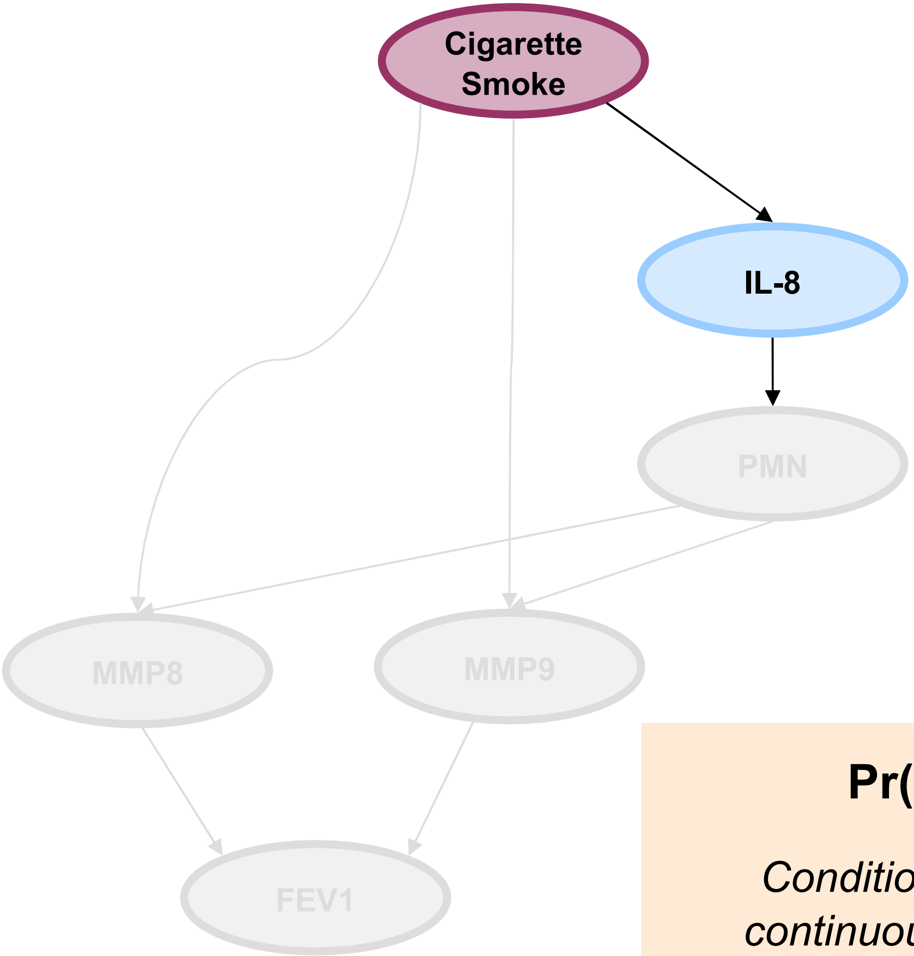
- Sputum samples
  - Whenever possible, use only data from sputum samples
  - Avoid standardization problems (e.g., concentration levels in BALF are quite different from concentration in sputum)
  - Clinical studies are more likely to use sputum samples (less invasive than BALF, etc.)

Case Study – Exclusion Criteria

- Data are excluded when…
  - … data refer to a group with a particular disease (e.g., asthma, cystic fibrosis, α1-AT deficiency, …)
  - … data have a unit that cannot be transformed to the same unit
  - … data are not biologically relevant
  - … smoking status is unknown or mixed when stratification by smoking status is required.

Case Study 1 (1/5)

Link Cigarette Smoke → IL-8



Pr(IL-8 | Cigarette Smoke) ???

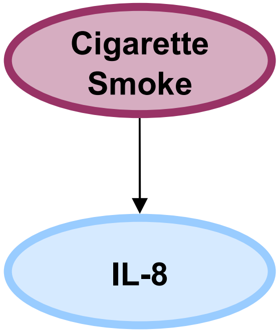
Conditional Probability Distribution (CPD) of a continuous variable given a categorical variable

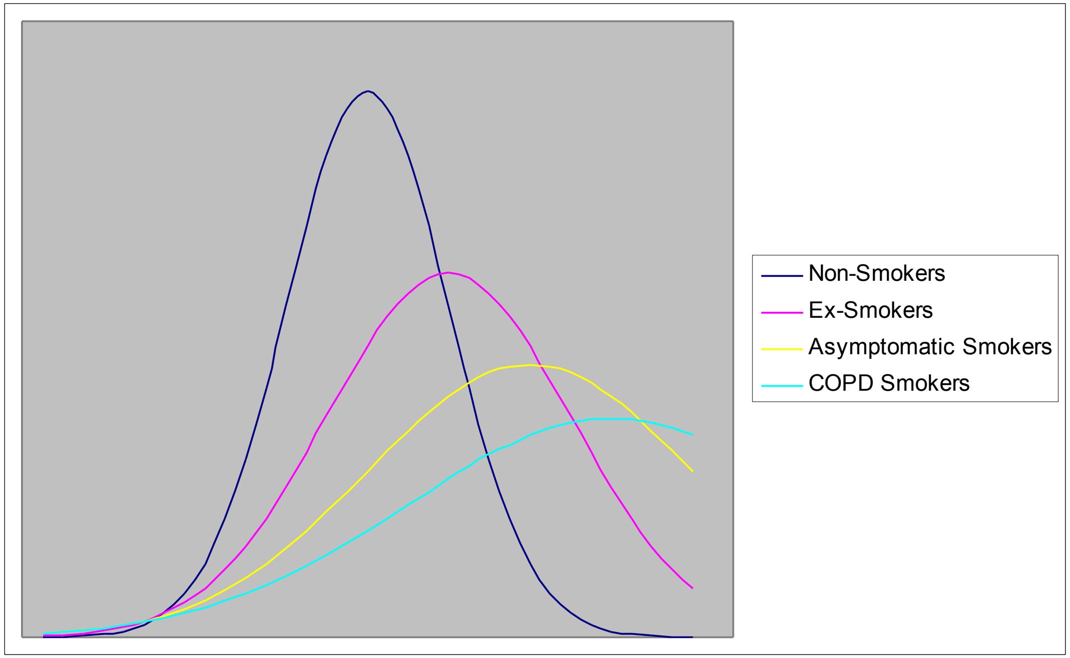
Case Study 1 (2/5)

Link Cigarette Smoke → IL-8

➢ Objective is to determine a probability distribution for each of the following categories…

- Non-Smokers
- Ex-Smokers
- Smokers without COPD
- Smokers with COPD





Case Study 1 (3/5)

Link Cigarette Smoke → IL-8

- Available data are rather limited
  - Usually low sample sizes
  - Frequently only aggregated data (means, SD, …)
  - Study context not always clearly specified
  - Outliers
- Data are heterogeneous
  - Very different populations: Age, gender, ethnic origin …

- However, a detailed analysis of the data is necessary…
  - … to include/exclude data
  - … to determine which data can be aggregated
  - … to decide whether an additional node (age, smoke-dose, …) needs to be added
- These points are also valid for all the other case studies!

Case Study 1 (4/5)

Link Cigarette Smoke → IL-8

- Values vary greatly in the literature.

- However, these variations cannot be attributed to anyone of the following confounders:
  - Age groups
  - Gender
  - Year of publication
  - Country of study
  - Smoke dose
  - Spirometric measurements (*e.g.*, FEV1)

Case Study 1 (5/5)

Link Cigarette Smoke → IL-8

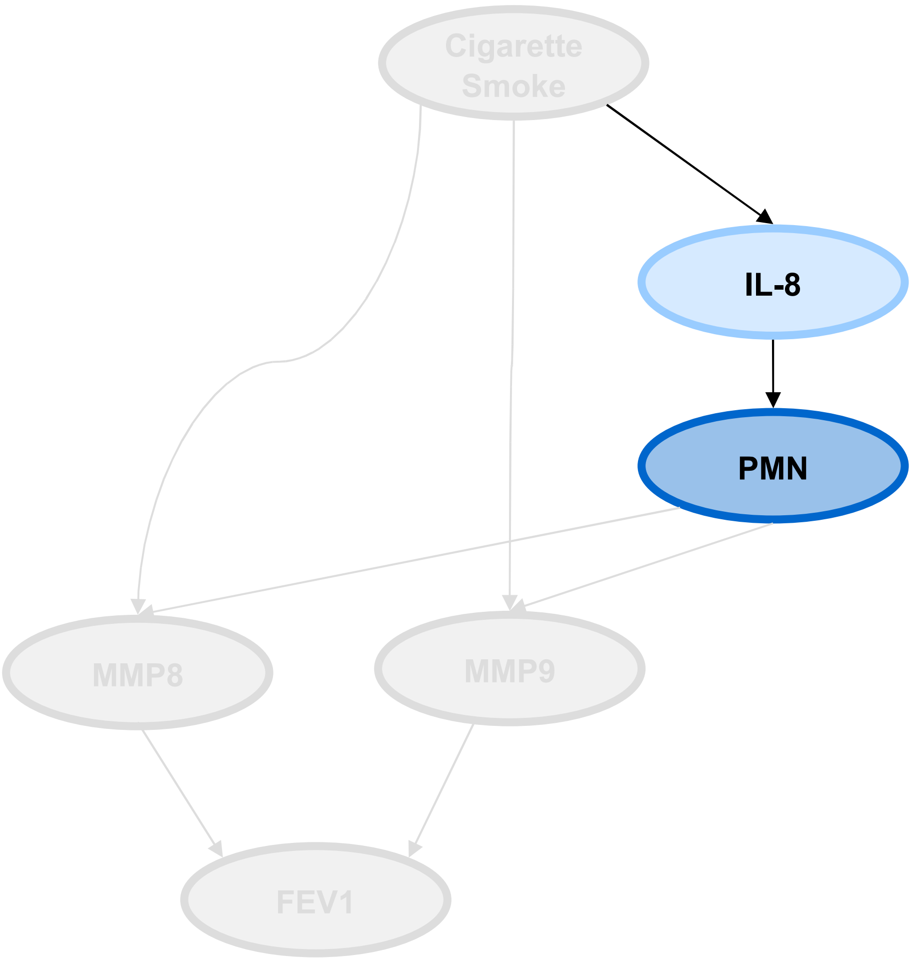
- We assume that the data are log normally distributed.

- For each observation, the parameters μ and σ of the underlying Log-Normal distribution are estimated using the summary statistics in the following order:
  - For μ:
    - Mean > Median > Geometric Mean > Interquartile Range > Range
  - For σ:
    - SD > SEM > Interquartile Range > Range

- Using ANOVA, we estimate an overall μ and σ, which represent the parameters of the CPD.

Case Study 2 (1/3)

Link IL-8 → PMN



Pr(PMN | IL-8) ???

CPD of a continuous variable given a continous variable

Case Study 2 (2/3)

Link IL-8 → PMN

- Data are not stratified by smoking status …
  - No link between *Cigarette Smoke* and *PMN* → Chemotactic activity of IL-8 towards neutrophils is *a priori* independent of the smoking status
    - First analysis will include all the data (independently of the smoking status)
    - In order to validate our assumption that chemotactic activity is independent of the smoking status, a second analysis will be performed on the data stratified by smoking status. These results will then be compared to the first analysis.
  - Same type of comment for healthy and COPD subjects
- Very difficult to find data where both variables (IL-8, PMN) are measured simultaneously.

Case Study 2 (3/3)

Link IL-8 → PMN

- In simple linear regression models, one assumes that the values of the regressor variable *x* are known constants, which is not the case for this example.

- The example shown here is a case where the regressor *x* is random. Therefore we treat the link of this type as a case where the two variables are jointly normally distributed:

$$f(y,x)=\frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}}\exp\left\{-\frac{1}{2(1-\rho^2)}\left[\left(\frac{y-\mu_1}{\sigma_1}\right)^2+\left(\frac{x-\mu_2}{\sigma_2}\right)^2-2\rho\left(\frac{y-\mu_1}{\sigma_1}\right)\left(\frac{x-\mu_2}{\sigma_2}\right)\right]\right\}$$

- The conditional distribution of *y* for a given value of *x* is:

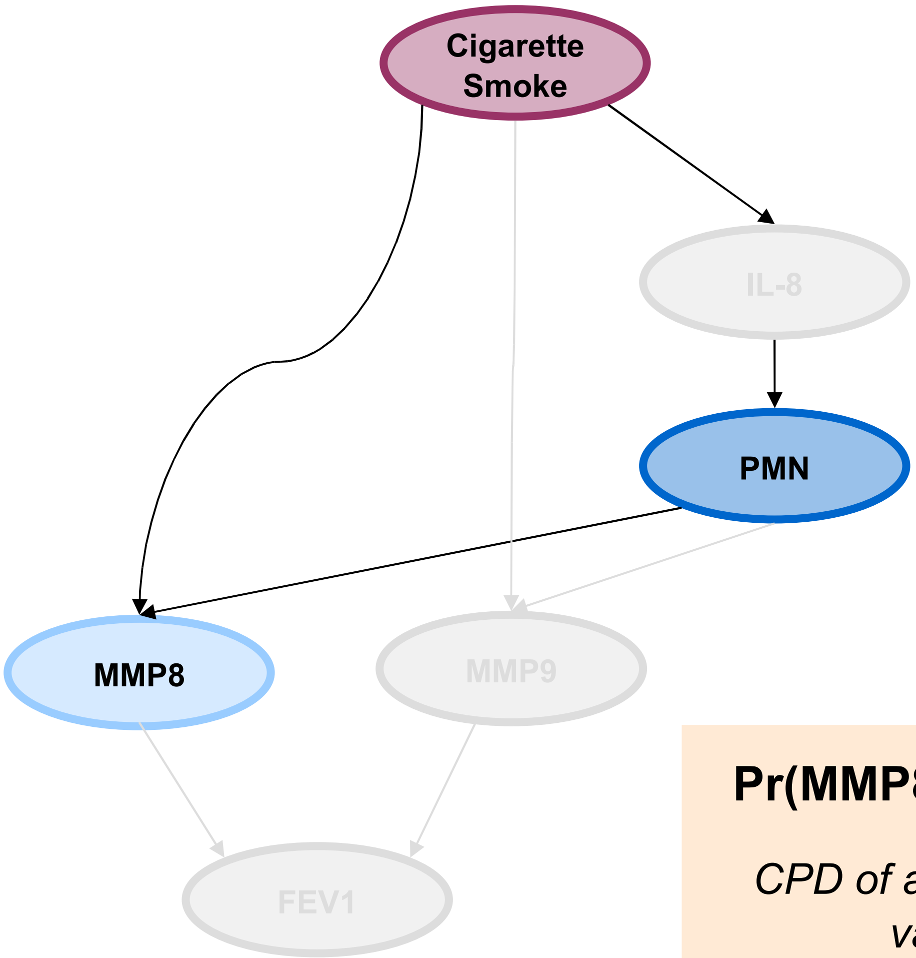
$$f(y|x)=\frac{1}{\sqrt{2\pi\sigma_{12}}}\exp\left[-\frac{1}{2}\left(\frac{y-\beta_0-\beta_1 x}{\sigma_{12}}\right)^2\right]$$

  - Where: , $\beta_1=\frac{\sigma_1}{\sigma_2}\rho$ and $\sigma_{1.2}=\sigma_1^2(1-\rho^2)$

- This conditional distribution would become the CPD of PMN.

Case Study 3 (1/2)

Link Cigarette Smoke + PMN → MMP8



Pr(MMP8 | Cigarette Smoke and PMN) ???

CPD of a continuous variable given a categorical variable and a continuous variable

Case Study 3 (1/2)

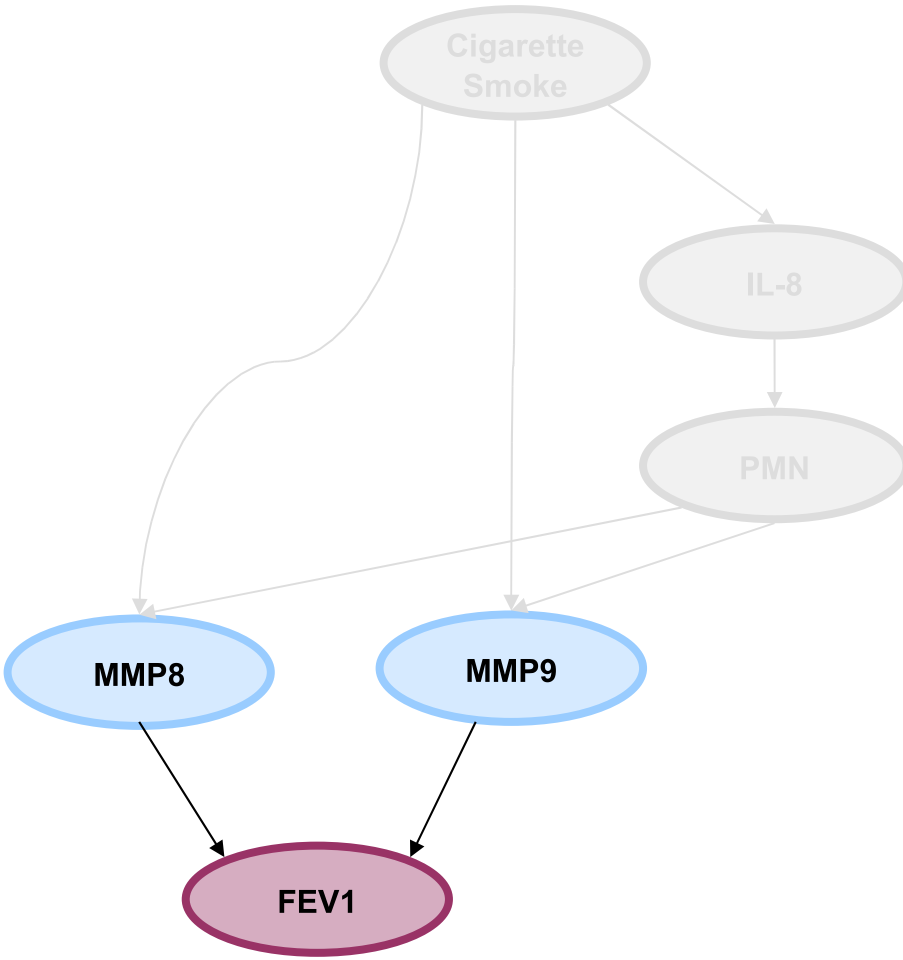
Link Cigarette Smoke + PMN → MMP8/MMP9

- Data are stratified by smoking status…
  - Link PMN → MMP8/MMP9 must be quantified for all categories of *Cigarette Smoke*

- Data analysis is similar to the case IL-8 → PMN, but it must be repeated for each category of *Cigarette Smoke* as for case study 1.

Case Study 4 (1/7)

Link MMP8 + MMP9 → FEV1



Pr(FEV1 | MMP8 and MMP9) ???

CPD of a continuous variable given (at least) 2 continuous variables

Case Study 4 (2/2)

Link MMP8 + MMP9 → FEV1

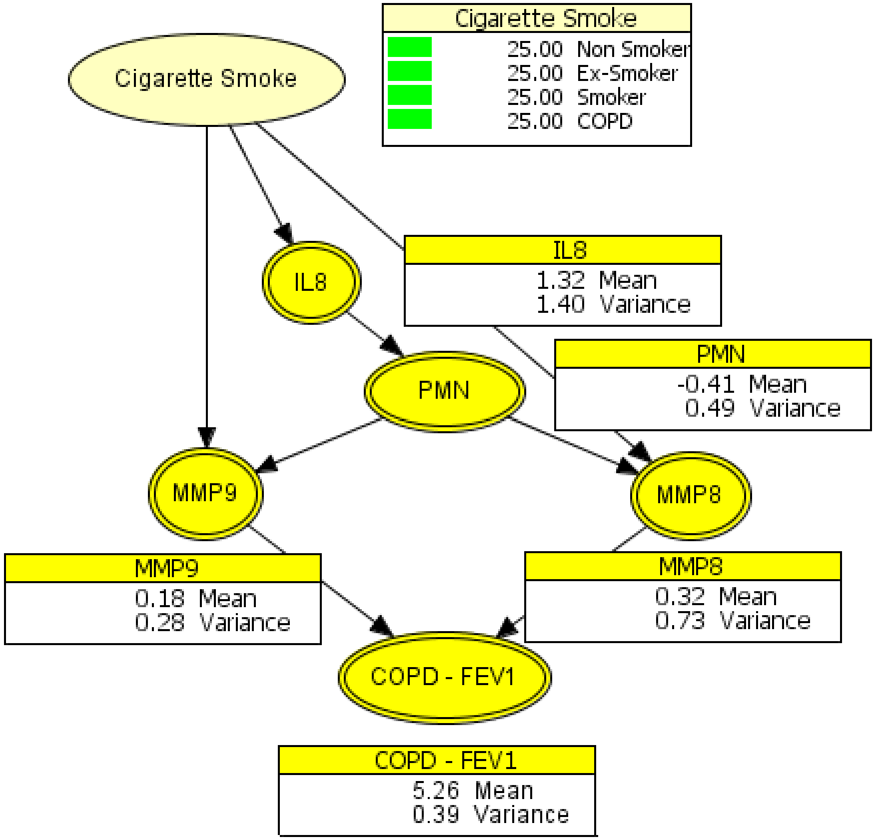
- Ideally, a regression equation of the type FEV1 = f (MMP8, MMP9) + ε and associated standard residual error is desired.

- Very difficult to find data where all 3 variables (MMP8, MMP9, FEV1) are measured simultaneously.

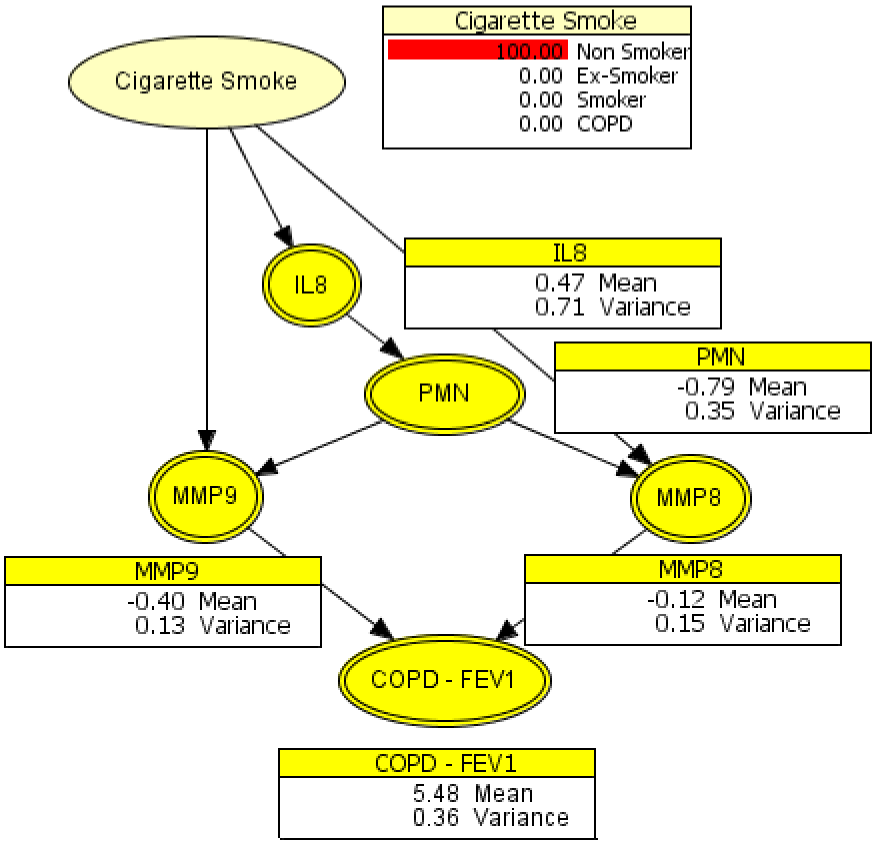
- Solution could be…
  - Treat the link MMP8 → FEV1 alone (as for the case IL-8 → PMN)
  - Treat the link MMP9 → FEV1 alone (as for the case IL-8 → PMN)
  - Compute a weighted average (weights could be chosen according to biological considerations)

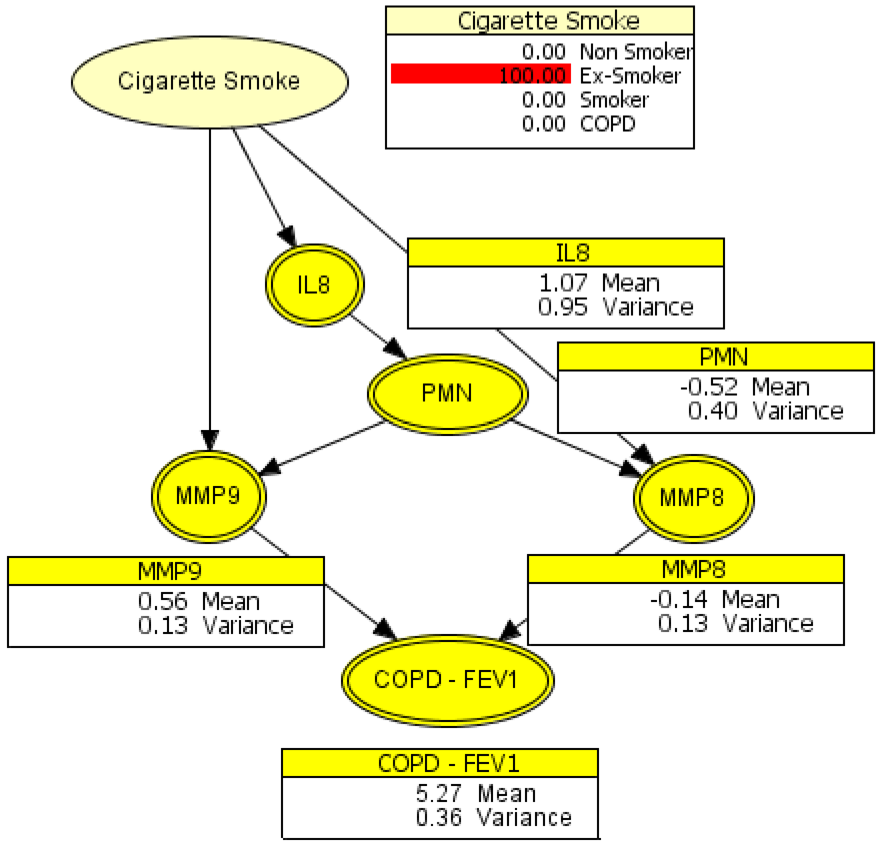
Hugin Model

- Population of 25% each for Non-Smokers, Ex-Smokers, asymptomatic Smokers, and Smokers with COPD:



Hugin Model

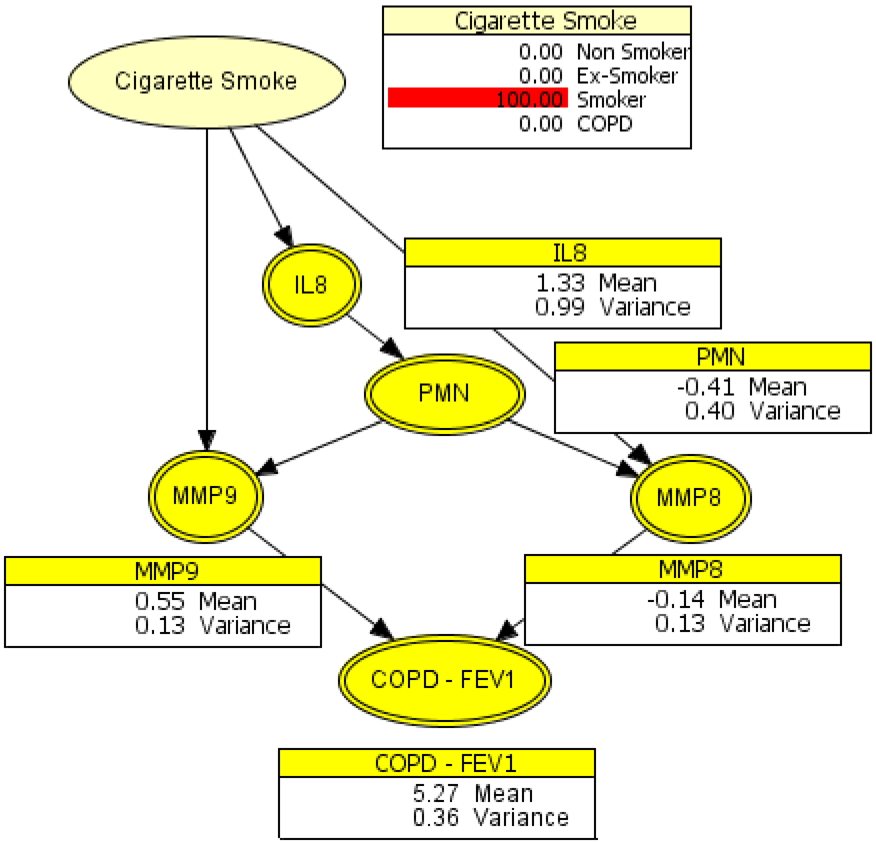


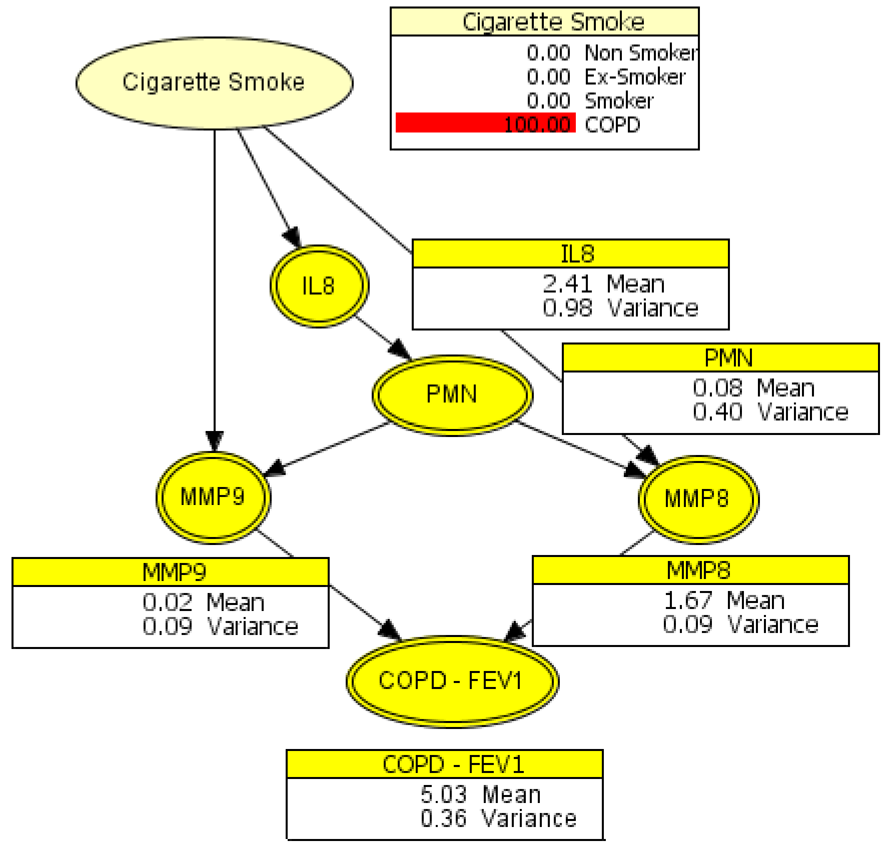


Population of 100% Non-Smokers

Population of 100% Ex-Smokers

Hugin Model





Population of 100% asymptomatic Smokers

Population of 100% Smokers with COPD

Issues

- Remaining issues:
  - Data gaps
    - Situation will improve with more articles in the database
    - New experiments will be required to fill these gaps
  - Data selection
    - Input from COPD experts is of prime importance
  - Mathematical analysis
    - Select better estimates for log-normal parameters
    - Test different distributions

- Data analysis may reveal patterns that are not currently represented in the model.

Next Steps

- Complete the human model
- Build an animal model
- Model Analysis & Validation
  - Conflict Analysis
  - Value of Information Analysis
  - Sensitivity Analysis
  - Cross-Validation on new data not used in the process of building the model

Acknowledgments

Pascal Cosandier

Christelle Haziza

Patrice Leroy

Carole Mathis

Michael J. Peck

Zheng Sponsiello Wang

Grégory Vuillaume

Rolf Weitkunat

Andrea Wohlsen

PMI