



UNC  
GILLINGS SCHOOL OF  
GLOBAL PUBLIC HEALTH

# **Machine Learning in Dose-Response Assessment**

## **Translating Science to Decisions**

**Jacqueline MacDonald Gibson, Associate Professor  
Gillings School of Global Public Health  
University of North Carolina at Chapel Hill**

**March 6, 2018**

# Outline

- **Introduction**
  - Current dose-response assessment methods: limitations
  - Bayesian belief networks as an alternative
- **Bayesian belief networks: background**
- **Example application: arsenic dose-response assessment**
  - Data source and machine learning method
  - Results
    - Predictive capability
    - Comparison to current approaches
  - Example policy application

# Currently Used Methods for Dose-Response Assessment

## Cancer

$$P(cancer) = \alpha$$



“slope factor”

# Currently Used Methods for Dose-Response Assessment

## Cancer

$$P(cancer) = \alpha$$



“slope factor”

## Other illnesses

$$Hazard = \frac{Dose}{RfD}$$



“reference dose”

# Current Approaches Have Limitations

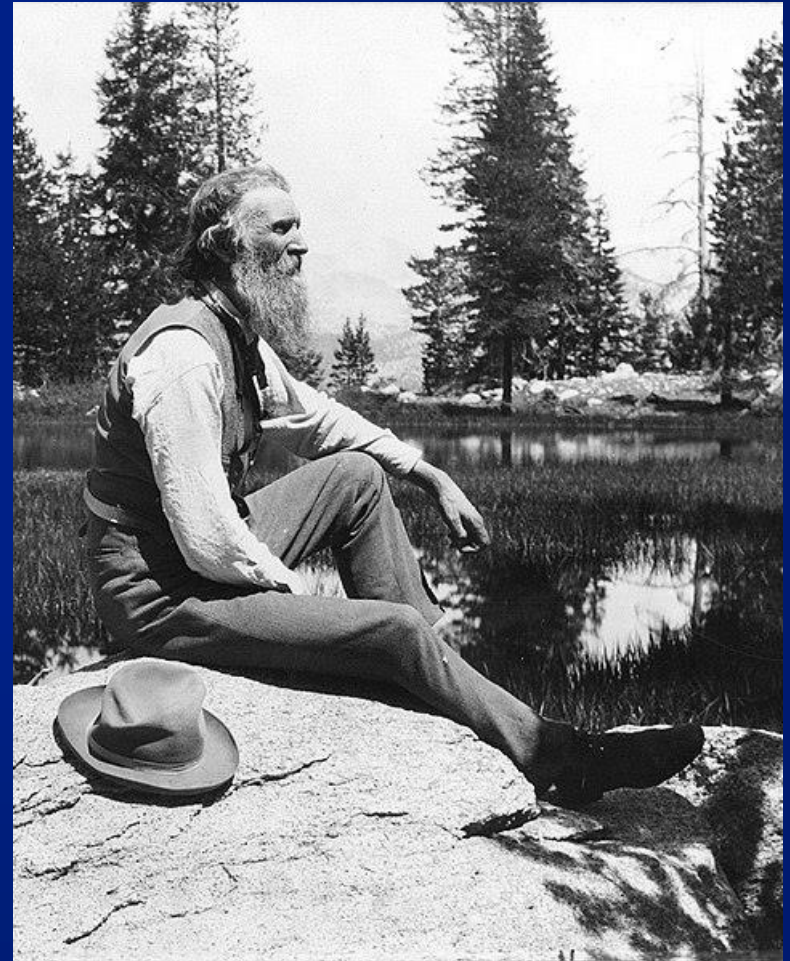
- **Cancer and noncancer methods differ**
  - Noncancer not quantified
- **Not customizable**
  - Generic slope factor applied to all
- **Nonlinear relationships not captured**
- **Risk factor interactions not considered**
  - E.g., Genetics, environment

**“When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.”**

**John Muir, 1869**

**Naturalist**

**Sierra Club Founder**



# Example: Arsenic Regulatory Impact Analysis

	Proposed Drinking Water Standard (µg/liter)			
	3	5	10	20
Net Benefits	\$(538.9)	\$ (287.4)	\$(111.2)	\$(31.8)
Benefit/Cost Ratio	0.16	0.24	0.32	0.48

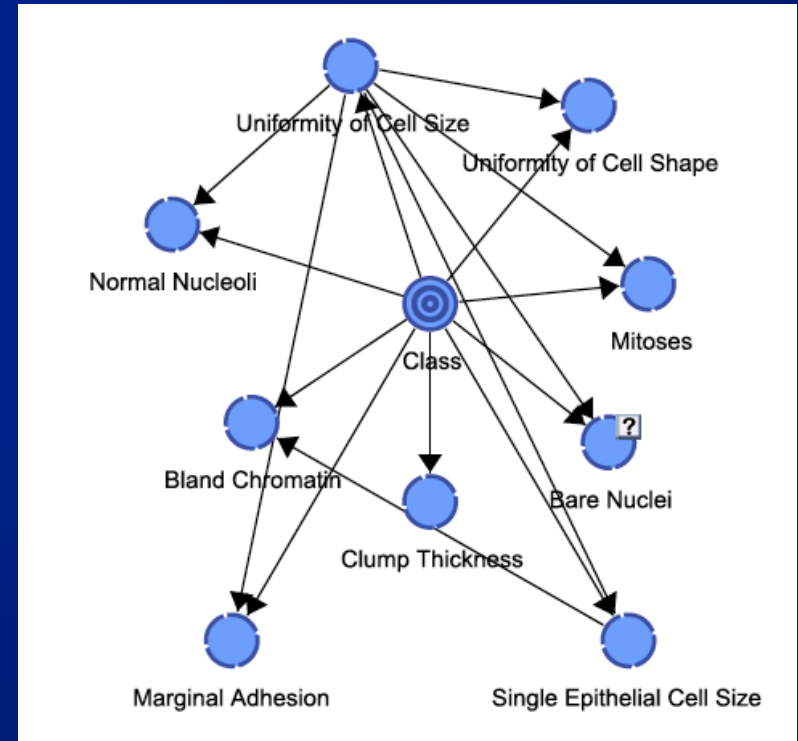
Proposed  
maximum  
contaminant  
level

Final  
maximum  
contaminant  
level

**Benefits** = avoided bladder cancer cases  
 =  $\alpha \times (\Delta Dose) \times Population$

# Bayesian Networks As Solution?

- Can include complex interactions
- Not restricted to linear or quasi-linear relationships
- Capability for individualized risk prediction
  - Personalized medicine analog



Breast cancer diagnosis  
example (Conrady and Jouffe,  
2011)

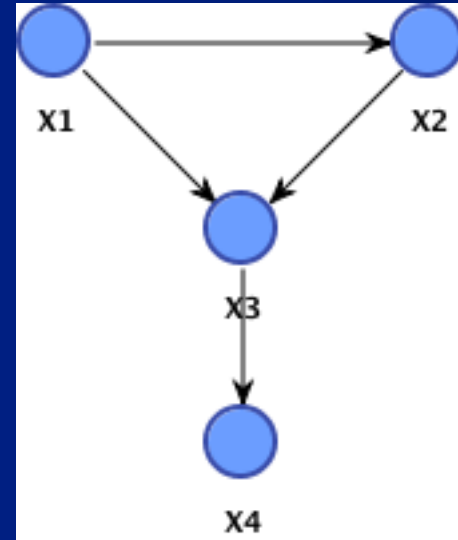


# Bayesian belief networks: background

# Bayesian Network Has Two Parts

## 1. Directed acyclic graph

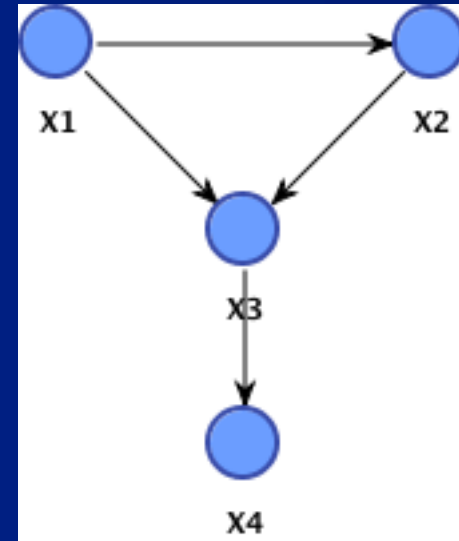
- Nodes=variables of interest
- Edges=relationships



# Bayesian Network Has Two Parts

## 1. Directed acyclic graph

- Nodes=variables of interest
- Edges=relationships

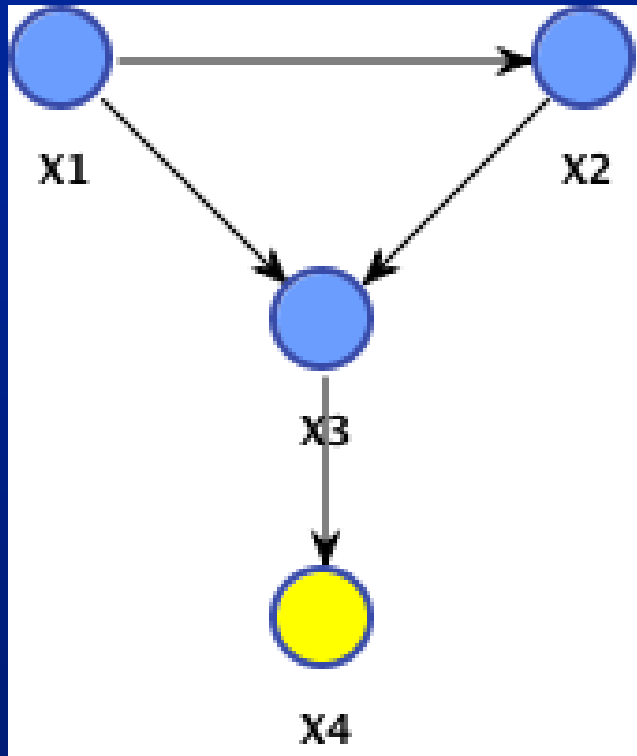


## 2. Joint probability distribution over the variables

- Conditional probability tables

X1	X2	False	True
False	False	10.000	90.000
	True	5.000	95.000
True	False	25.000	75.000
	True	10.000	90.000

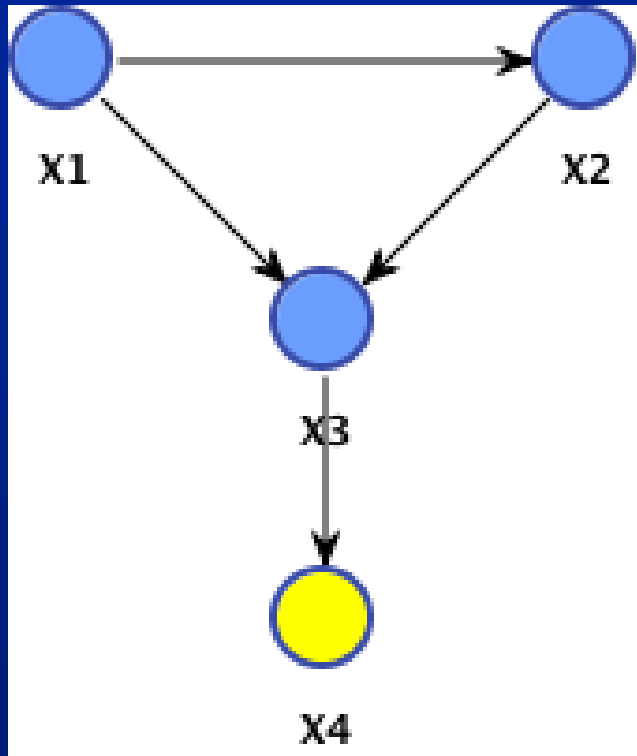
# Bayes' Theorem Used to Update Nodes with Evidence



$$P(X3|X4)$$

$$= \frac{P(X4|X3) \times P(X3)}{P(X4)}$$

# Bayes' Theorem Used to Update Nodes with Evidence



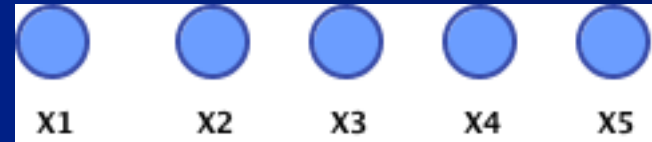
$$P(X3|X4)$$

$$= \frac{P(X4|X3) \times P(X3)}{P(X4)}$$

**All Bayesian methods are not equivalent.**

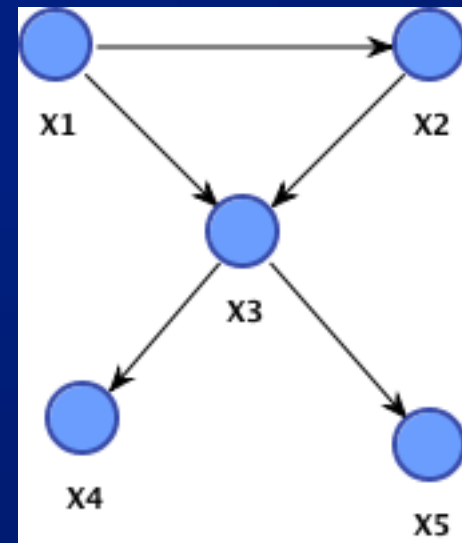
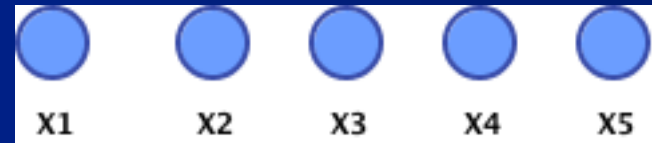
# Originated in Artificial Intelligence

- Early AI challenge: compact representation of data
  - Need  $2^5 - 1 = 31$  parameters to represent joint distribution



# Originated in Artificial Intelligence

- Early AI challenge: compact representation of data
  - Need  $2^5 - 1 = 31$  parameters to represent joint distribution
- Compact representation via conditional independencies
  - 17 parameters instead of 31



# First Solution Algorithms in Late 1980s

## A Turing Award for Helping Make Computers Smarter

BY STEVE LOHR MARCH 15, 2012 3:00 AM 14

Email

Share

Tweet

Save

More

Google search, I.B.M.'s Watson Jeopardy-winning computer, credit-card fraud detection and automated speech recognition.

There seems not much in common on that list. But it is a representative sampling of the kinds of modern computing chores that use the ideas and technology developed by Judea Pearl, the winner of this year's Turing Award.

The award, often considered the computer science equivalent of a Nobel prize, was announced on Wednesday by the [Association for Computing Machinery](#).

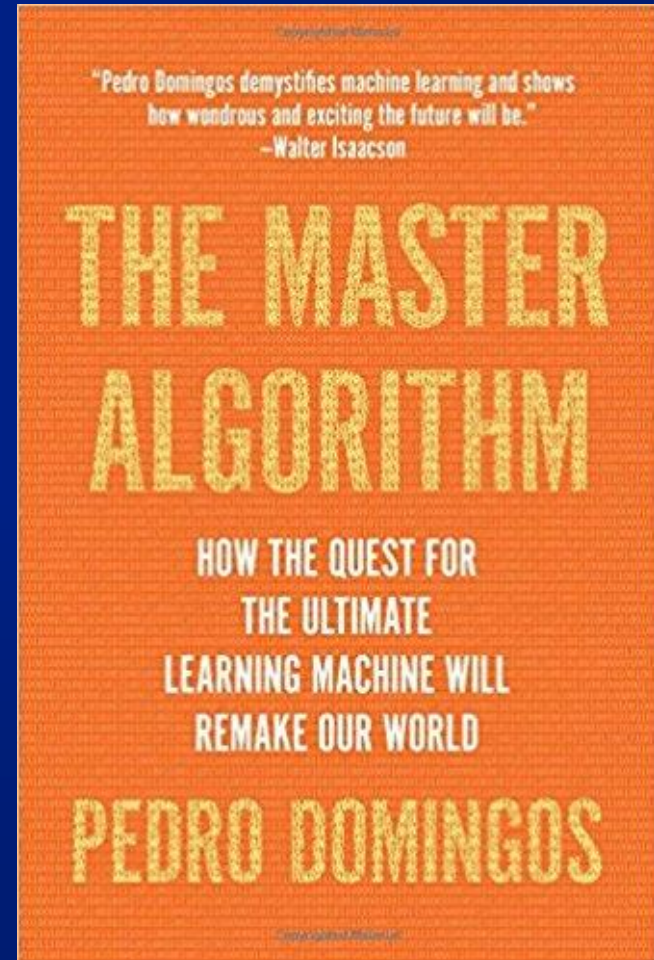


Judea Pearl, winner of the Turing Prize.



# Example Applications

- Spam filtering
- HIV vaccine development
- Infectious disease diagnosis
- Google AdSense
- Microsoft Xbox Live player rating



# **Example application: arsenic dose-response assessment**

# Arsenic Has Many Health Effects

- High doses long known to cause blackfoot disease
- Established associations with bladder, lung cancers
- Emerging evidence of association with diabetes



# Data from Mexican Cohort



Study area: Chihuahua, Mexico

- 1,050 adults  $\geq 18$  years old:
  - 880 without diabetes
  - 170 with diabetes

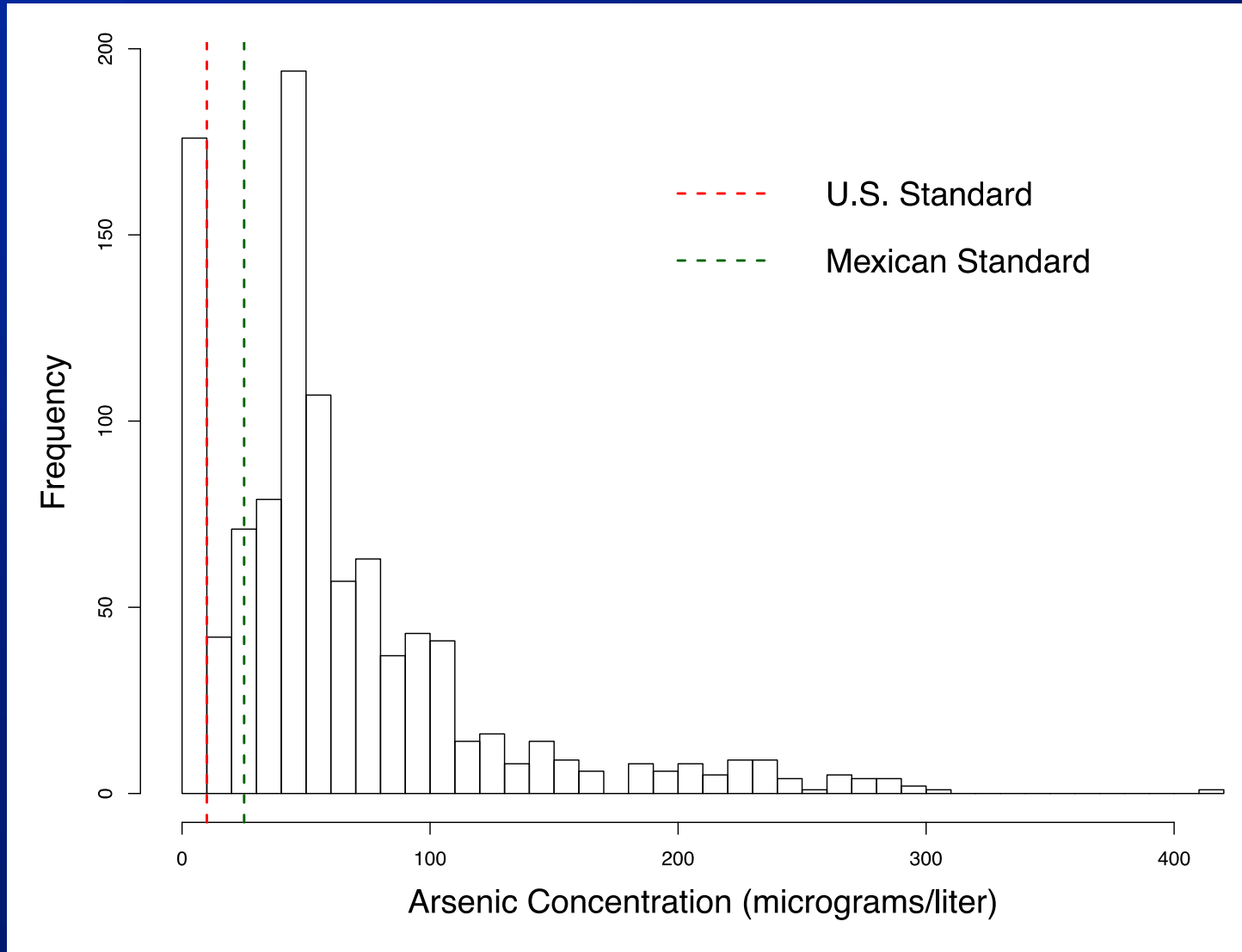
# Data from Mexican Cohort



Study area: Chihuahua, Mexico

- **1,050 adults  $\geq 18$  years old:**
  - 880 without diabetes
  - 170 with diabetes
- **Variables in data set:**
  - Arsenic in drinking water
  - Arsenic and metabolites in urine
  - Water source
  - Diet
  - Smoking
  - Anthropometry: BMI, waist size
  - Age, gender, education, ethnicity

# High Arsenic Exposure in Study Area



# Machine-Learned Network

- **Using BayesiaLab software:**
  1. Search “equivalence classes” of possible networks
  2. Keep nodes within five links of diabetes
  3. Eliminate nodes not significantly related to diabetes
  4. Re-run using augmented naïve Bayes algorithm
- **Test via five-fold cross-validation**

# Comparison to Traditional Approaches

## Reference Dose Method

$$\begin{aligned} &P(\text{diabetes}) \\ &= \begin{cases} 0 & \text{if } [As] \leq RfC \\ 1 & \text{if } [As] > RfC \end{cases} \end{aligned}$$



# Comparison to Traditional Approaches

## Reference Dose Method

$$\begin{aligned} &P(\text{diabetes}) \\ &= \begin{cases} 0 & \text{if } [As] \leq RfC \\ 1 & \text{if } [As] > RfC \end{cases} \end{aligned}$$

RfC (reference concentration)  
from EPA Integrated Risk  
Information System:

$$\begin{aligned} &3 \times 10^{-4} \frac{\text{mg}}{\text{kg-day}} \\ &\approx 10.5 \mu\text{g/l} \end{aligned}$$

# Comparison to Traditional Approaches

## Reference Dose Method

$$\begin{aligned} P(\text{diabetes}) \\ = \begin{cases} 0 & \text{if } [As] \leq RfC \\ 1 & \text{if } [As] > RfC \end{cases} \end{aligned}$$

## Slope Factor Method

$$\begin{aligned} P(\text{diabetes}) \\ = \text{slope factor} \times \text{Dose} \end{aligned}$$

RfC (reference concentration)  
from EPA Integrated Risk  
Information System:

$$\begin{aligned} 3 \times 10^{-4} \frac{\text{mg}}{\text{kg-day}} \\ \approx 10.5 \mu\text{g/l} \end{aligned}$$

# Comparison to Traditional Approaches

## Reference Dose Method

$$P(\text{diabetes}) = \begin{cases} 0 & \text{if } [As] \leq RfC \\ 1 & \text{if } [As] > RfC \end{cases}$$

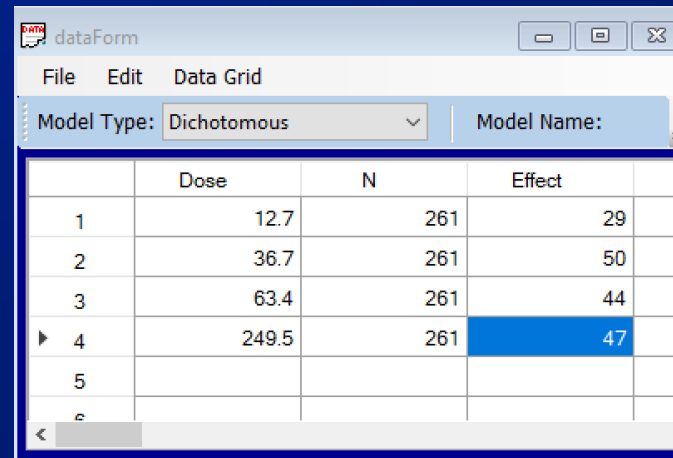
RfC (reference concentration)  
from EPA Integrated Risk  
Information System:

$$3 \times 10^{-4} \frac{\text{mg}}{\text{kg-day}} \\ \approx 10.5 \mu\text{g/l}$$

## Slope Factor Method

$$P(\text{diabetes}) = \text{slope factor} \times \text{Dose}$$

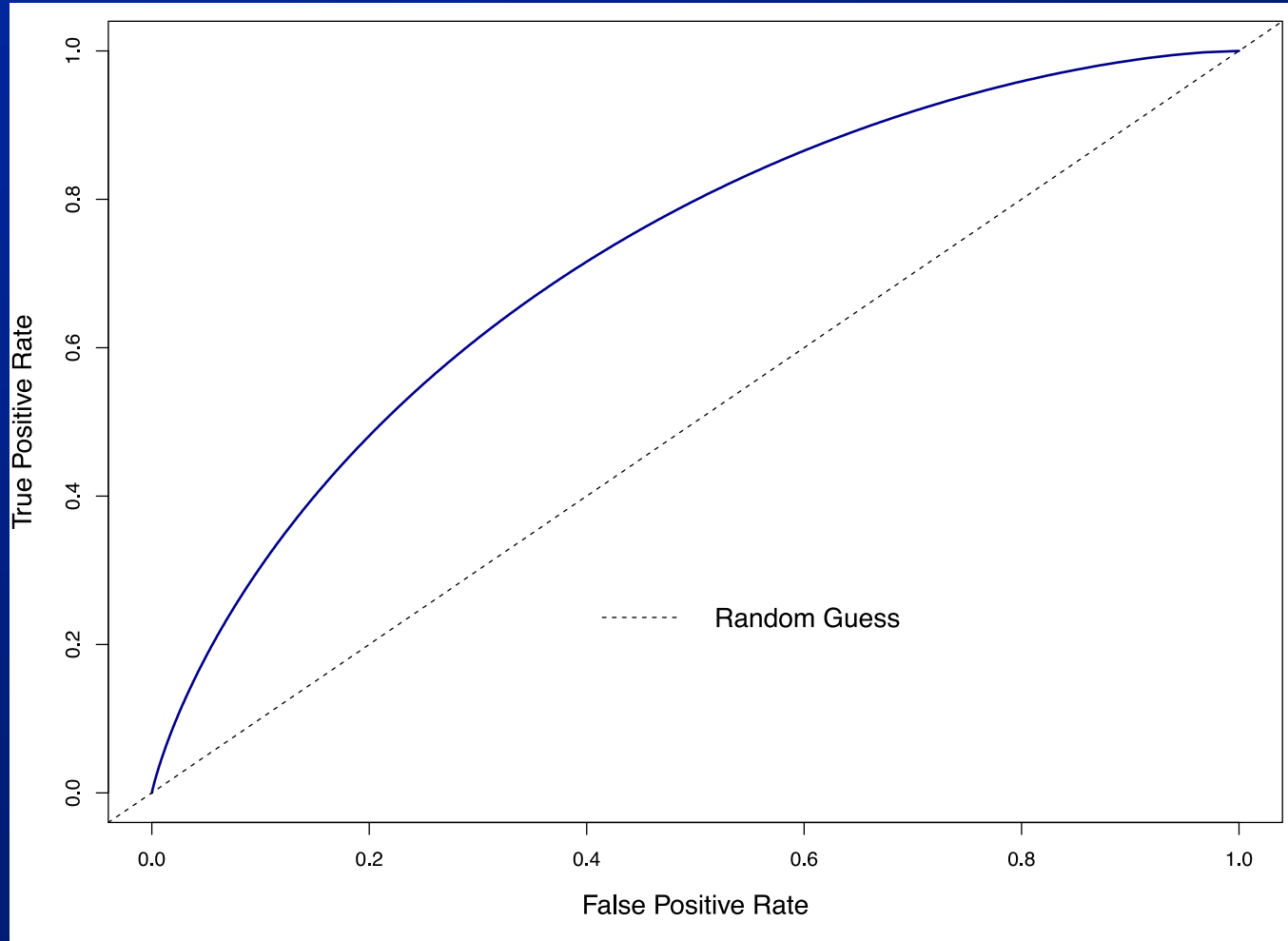
Slope factor estimated with  
Benchmark Dose Software



The screenshot shows the 'dataForm' window of Benchmark Dose Software. It has a menu bar with 'File', 'Edit', and 'Data Grid'. Below the menu bar, there are two fields: 'Model Type:' set to 'Dichotomous' and 'Model Name:'. Below these fields is a data grid with the following data:

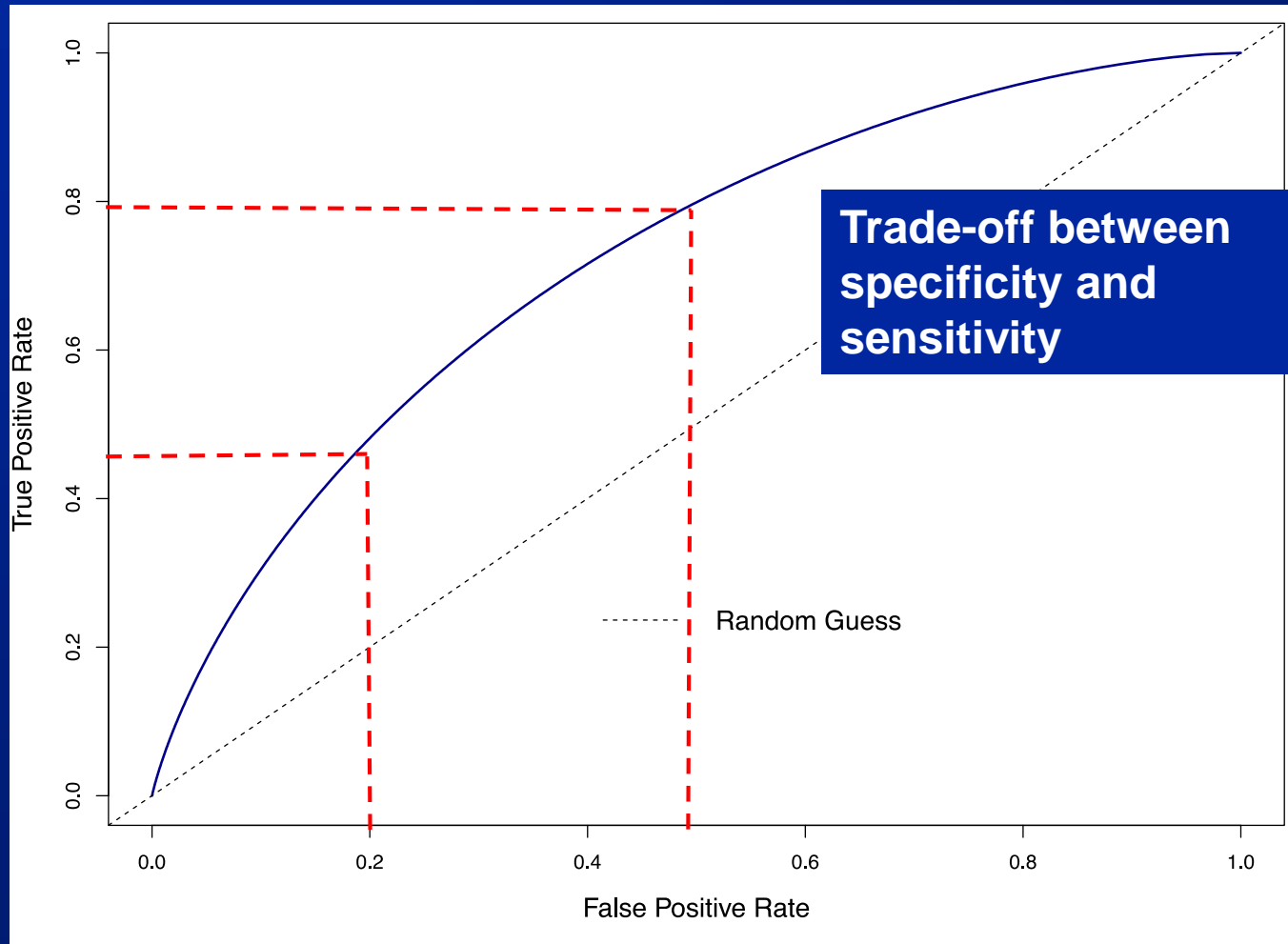
	Dose	N	Effect
1	12.7	261	29
2	36.7	261	50
3	63.4	261	44
4	249.5	261	47
5			
6			

# Compare Using Receiver-Operating Characteristic (ROC) Curves



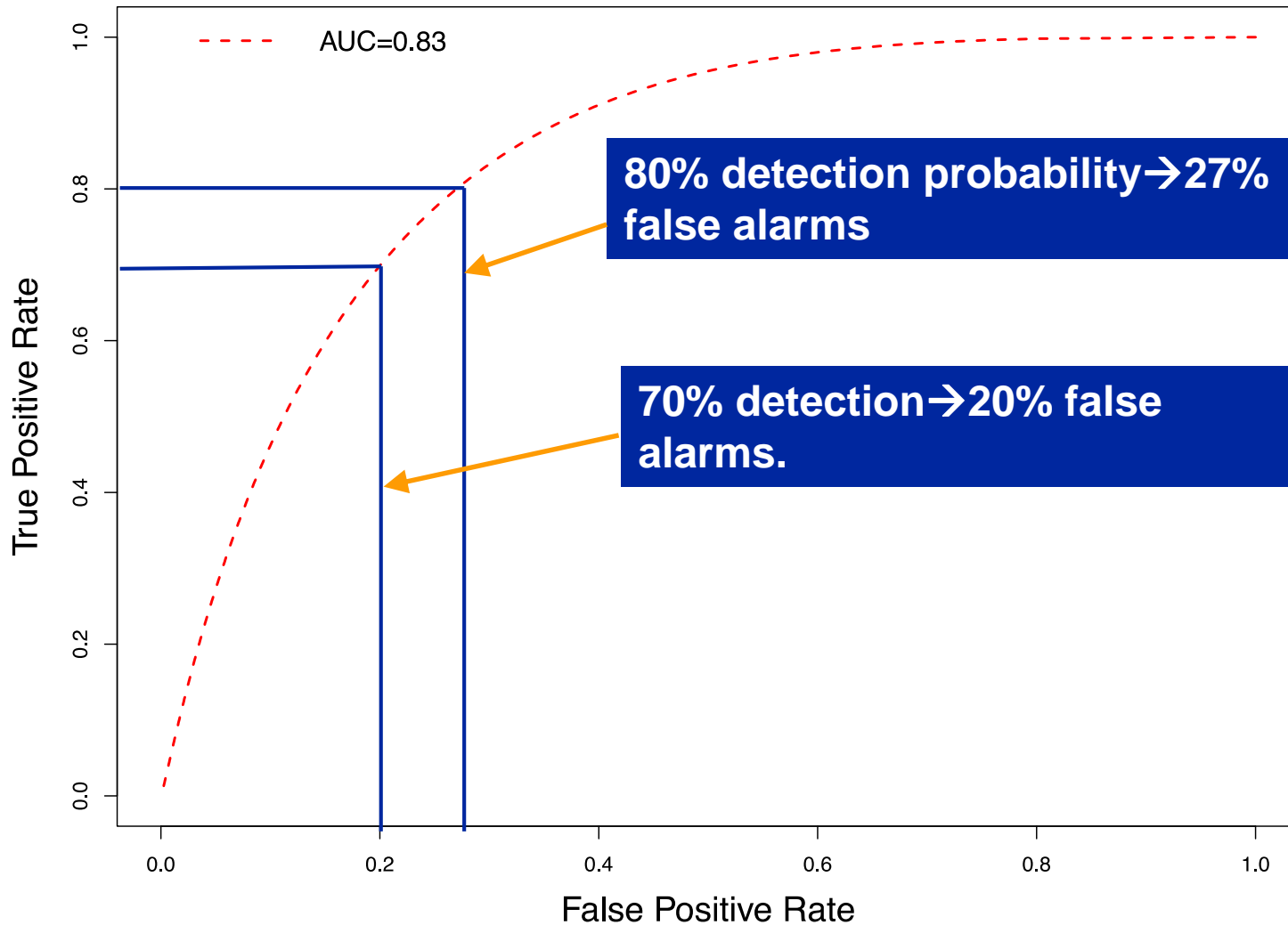
**Diagnostic accuracy = area under curve (1=perfect)**

# Compare Using Receiver-Operating Characteristic (ROC) Curves

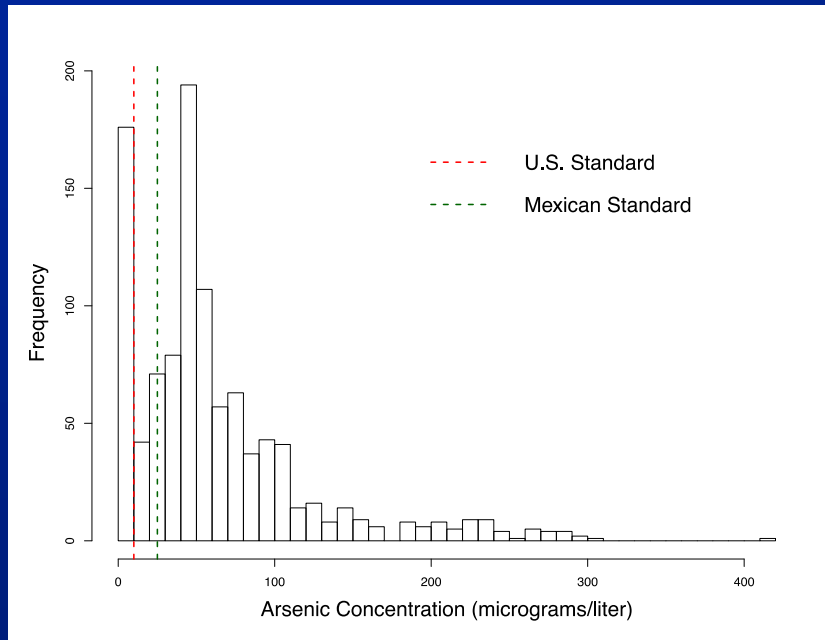


**Diagnostic accuracy = area under curve (1=perfect)**

# Interpretation: Trading Off Sensitivity, Specificity



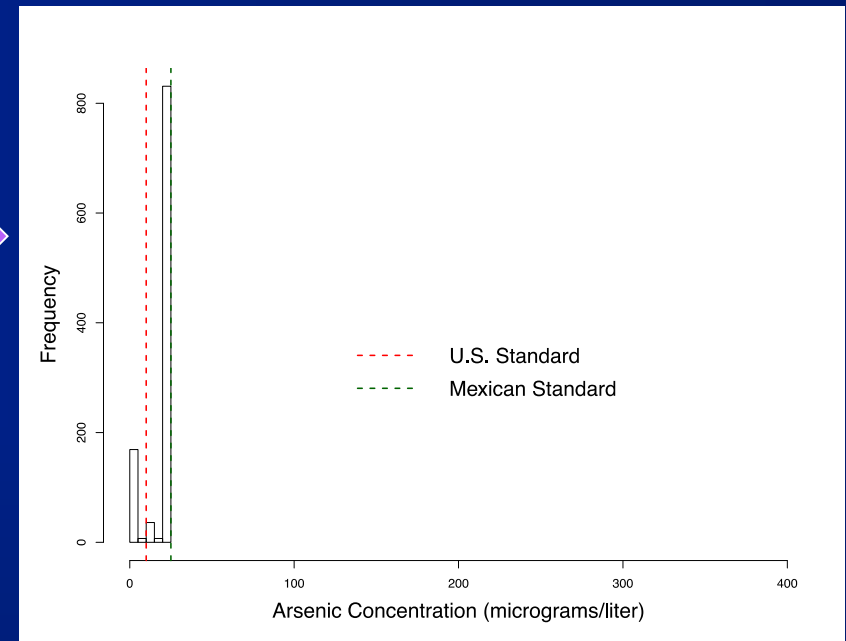
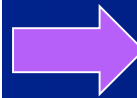
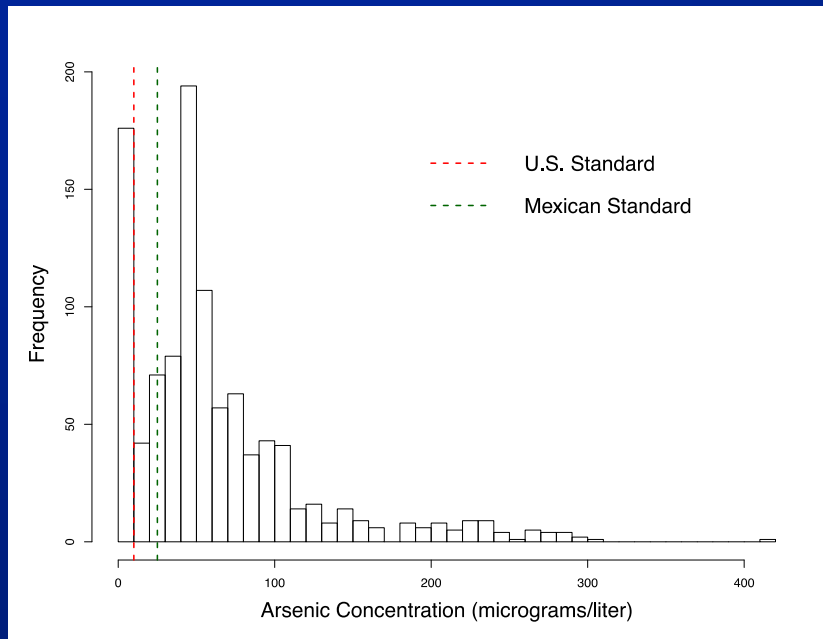
# Policy Analysis Example Considers Change in Risk If Decrease Exposure



**Treat all drinking water to  $< 25 \mu\text{g/liter}$**

- “Generic” population
- Vulnerable population: age $>50$ , metabolic risk factors

# Policy Analysis Example Considers Change in Risk If Decrease Exposure



**Treat all drinking water to  $< 25 \mu\text{g/liter}$**

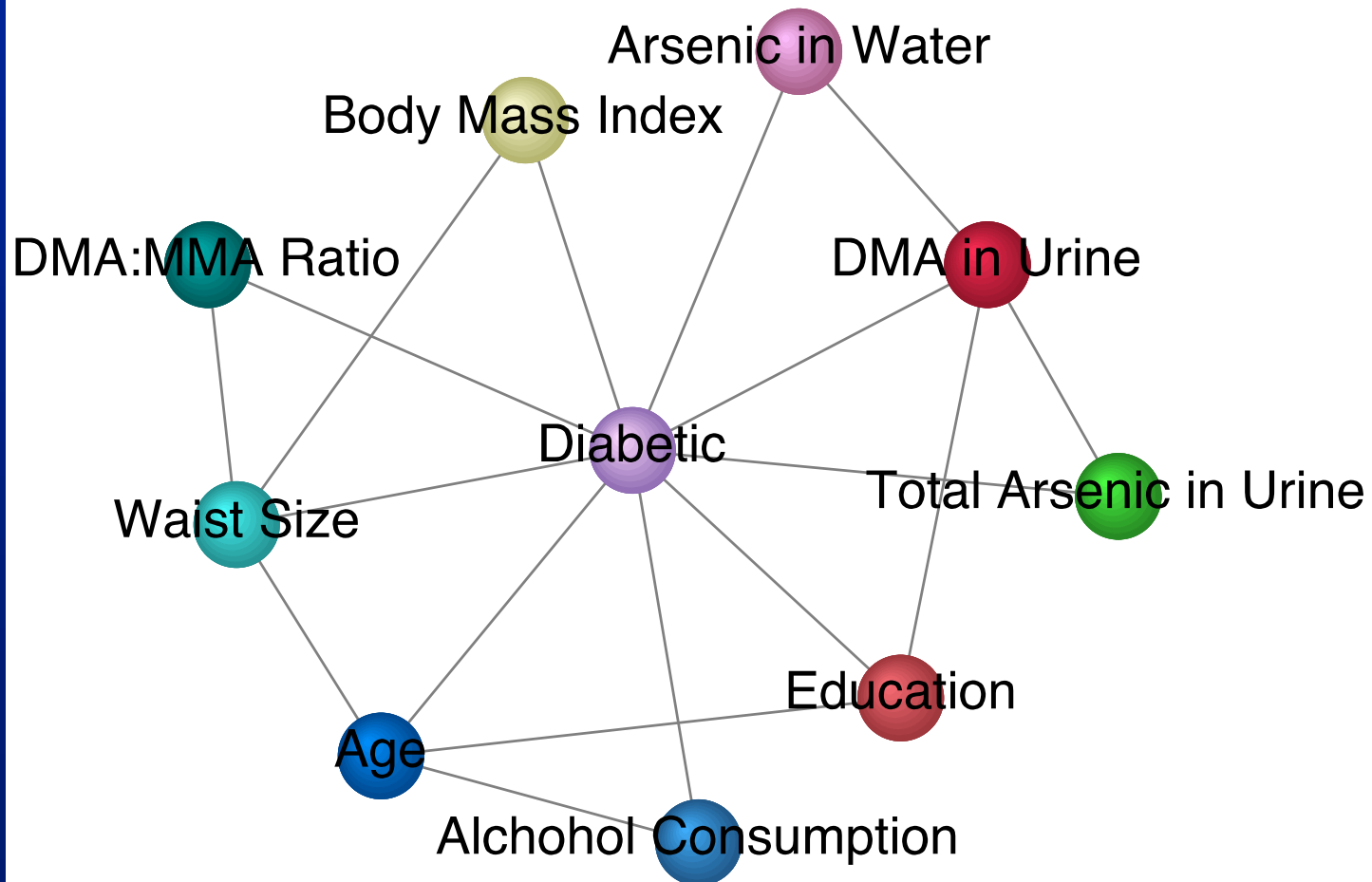
- “Generic” population
- Vulnerable population: age>50, metabolic risk factors



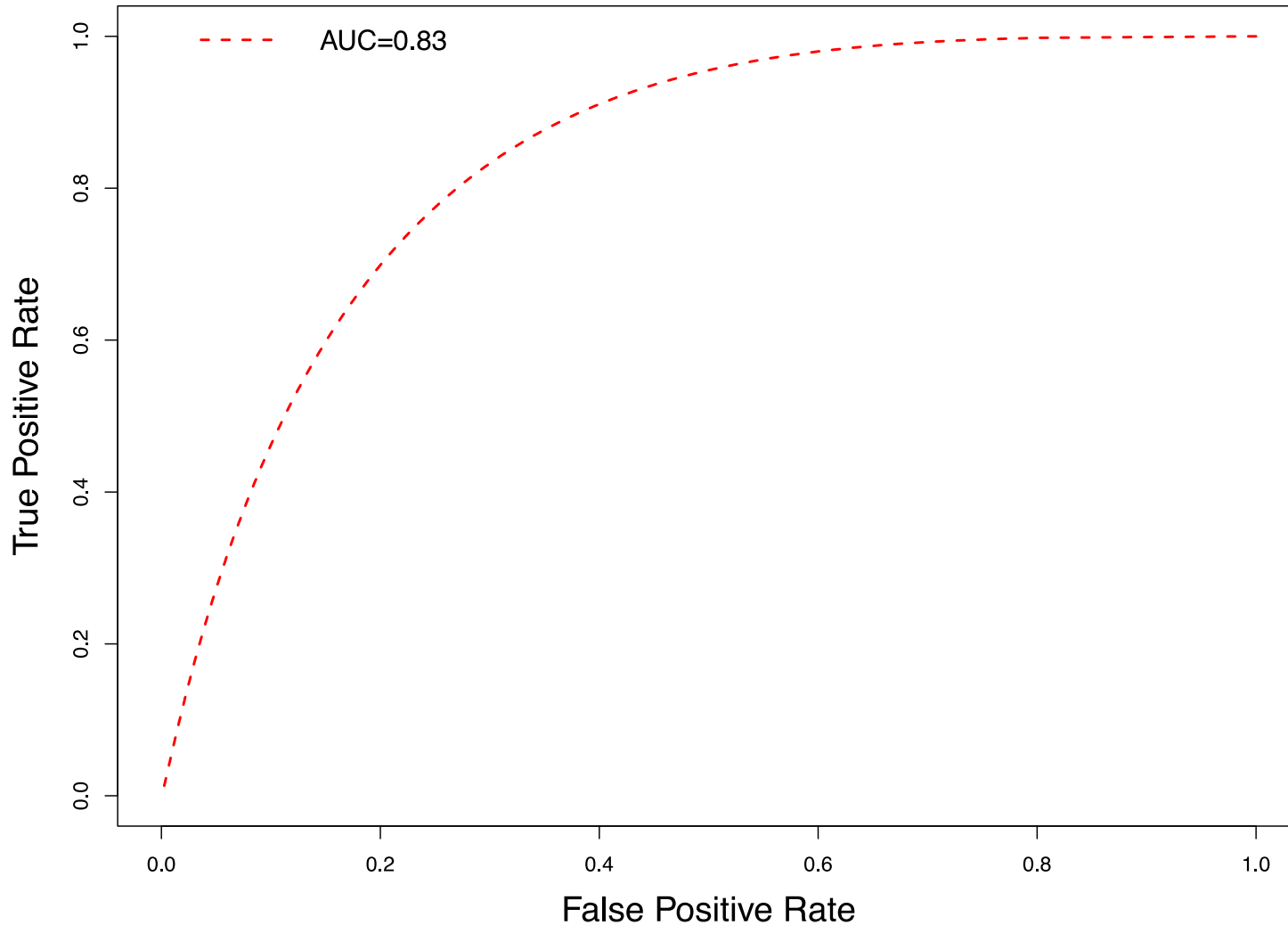
# Results

# Learned Network Shows Multiple Connections

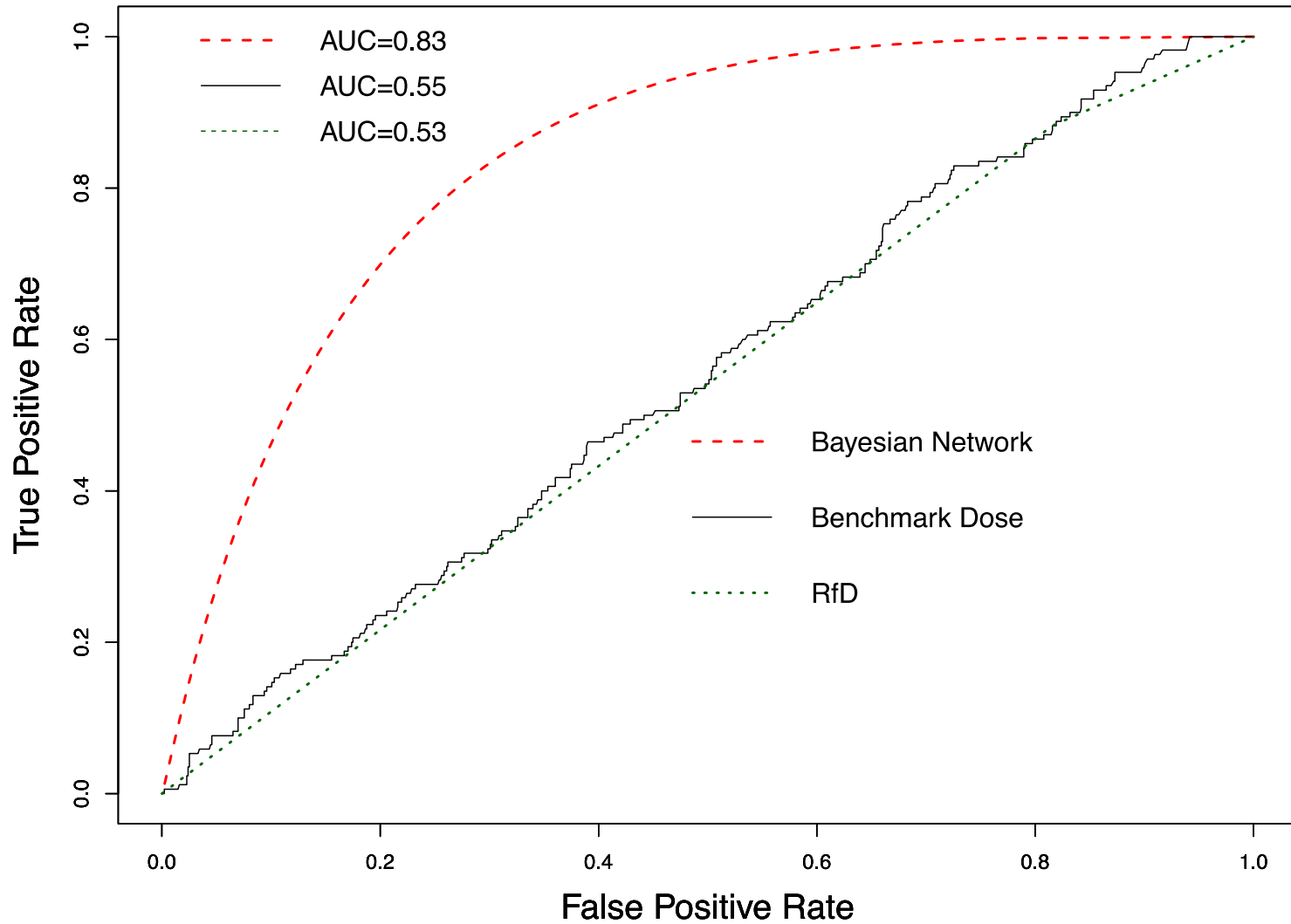
“Invisible cords that cannot be broken”



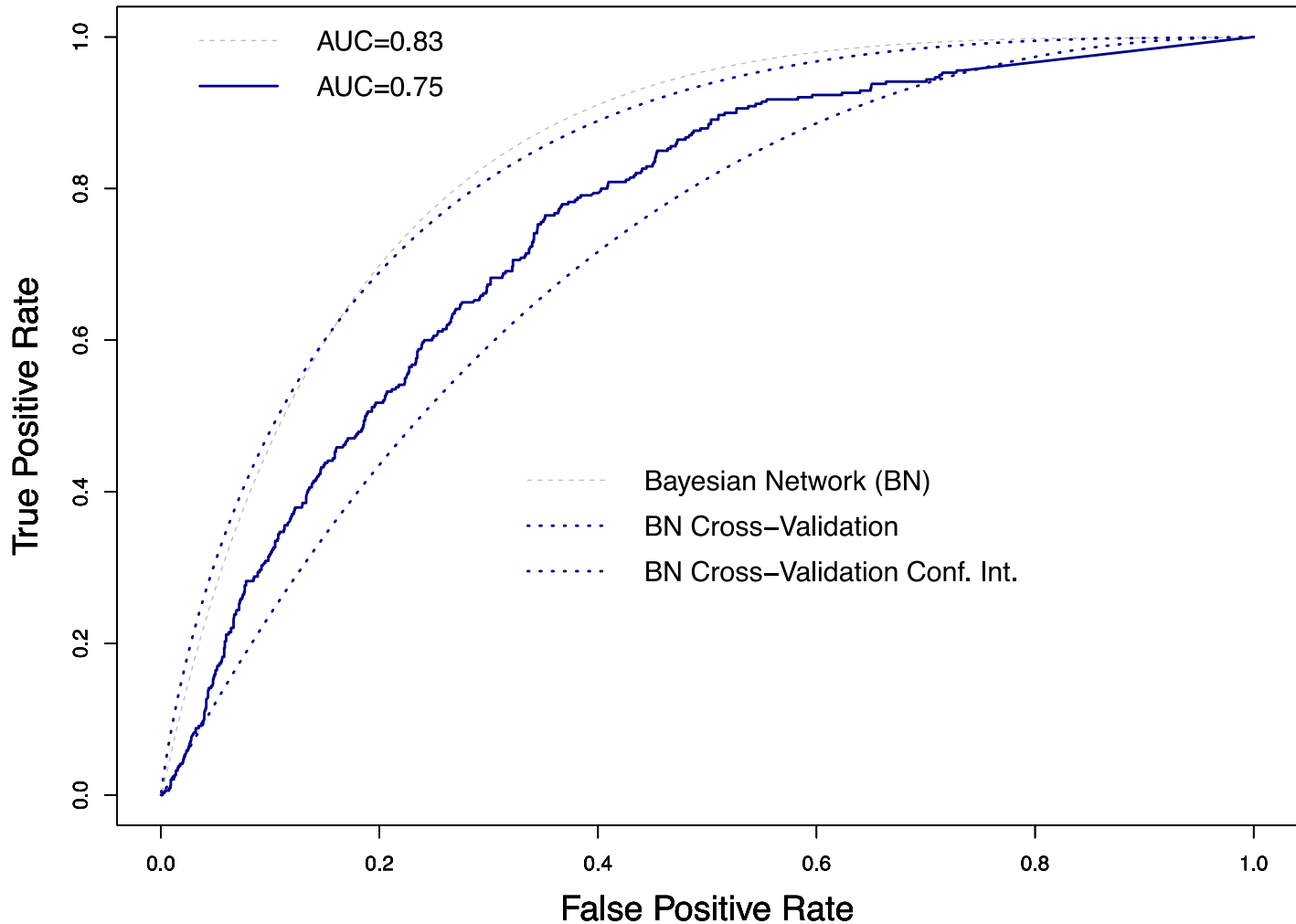
# Bayes Net Model Has High Predictive Ability



# Network Outperforms Current Methods



# Network Performance Well Maintained Under Cross-Validation



# Conventional Risk Analysis Overlooks Noncancer Benefits of Intervention

- Scenario: Arsenic < 25 µg/l in all water
- Cancer benefit:
$$\begin{aligned}\Delta\text{Cases} &= \text{slope factor} \times (\Delta\text{Concentration}) \times N \\ &= 0.0005 \times \Delta\text{Concentration} \times 1050 \\ &= 2\end{aligned}$$

# Conventional Risk Analysis Overlooks Noncancer Benefits of Intervention

- Scenario: Arsenic < 25 µg/l in all water

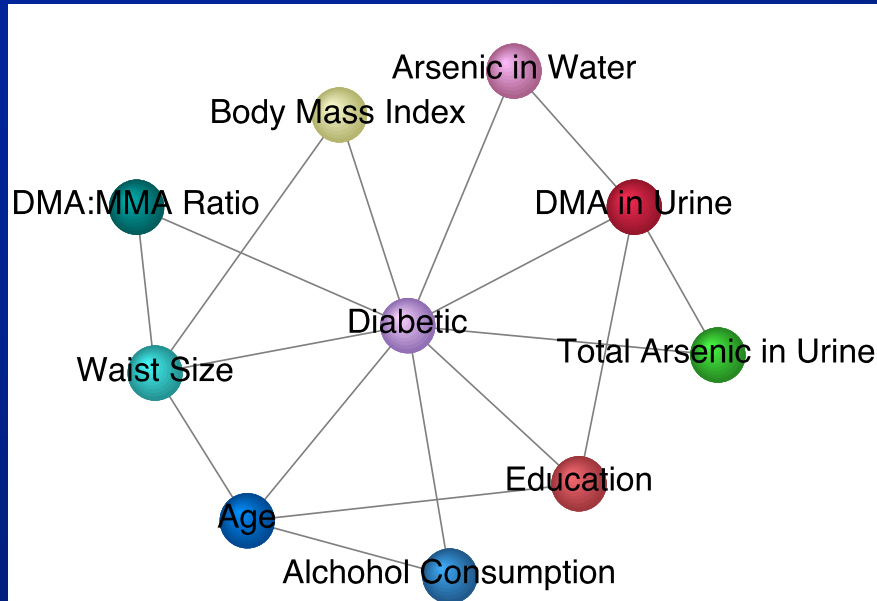
- Cancer benefit:

$$\begin{aligned}\Delta \text{Cases} &= \text{slope factor} \times (\Delta \text{Concentration}) \times N \\ &= 0.0005 \times \Delta \text{Concentration} \times 1050 \\ &= 2\end{aligned}$$

- Non-cancer benefit:

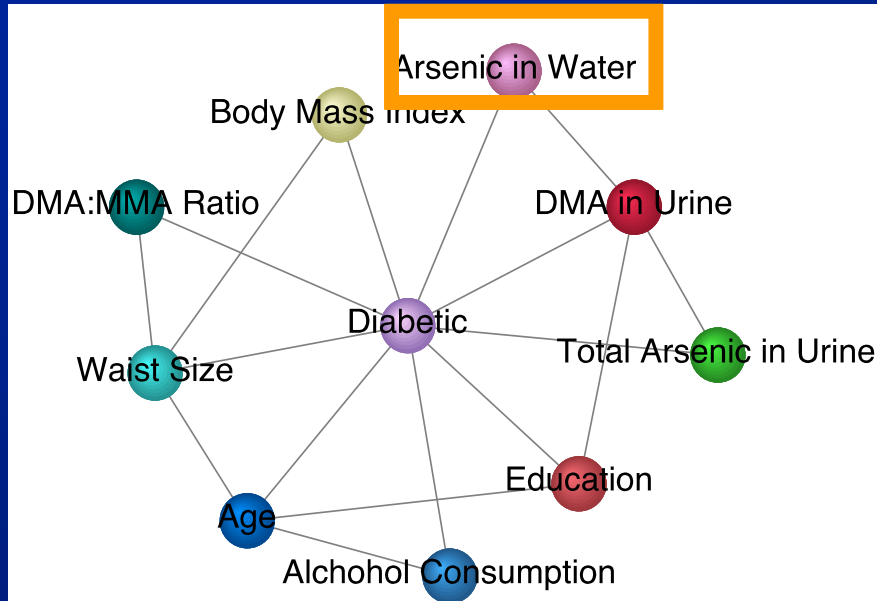
$$\begin{aligned}\Delta \text{At-risk population} &= \sum_{i=1}^{1050} I(HQ_0 > 1) - \sum_{i=1}^{1050} I(HQ_{\text{intervention}} > 1) \\ &= \sum_{i=1}^{1050} I\left(\frac{C_0}{10.5 \mu\frac{g}{l}} > 1\right) - \sum_{i=1}^{1050} I\left(\frac{C_{\text{intervention}}}{10.5 \mu\frac{g}{l}} > 1\right) \\ &= 0\end{aligned}$$

# Bayesian Network Can Estimate Diabetes Risk Reduction Benefits

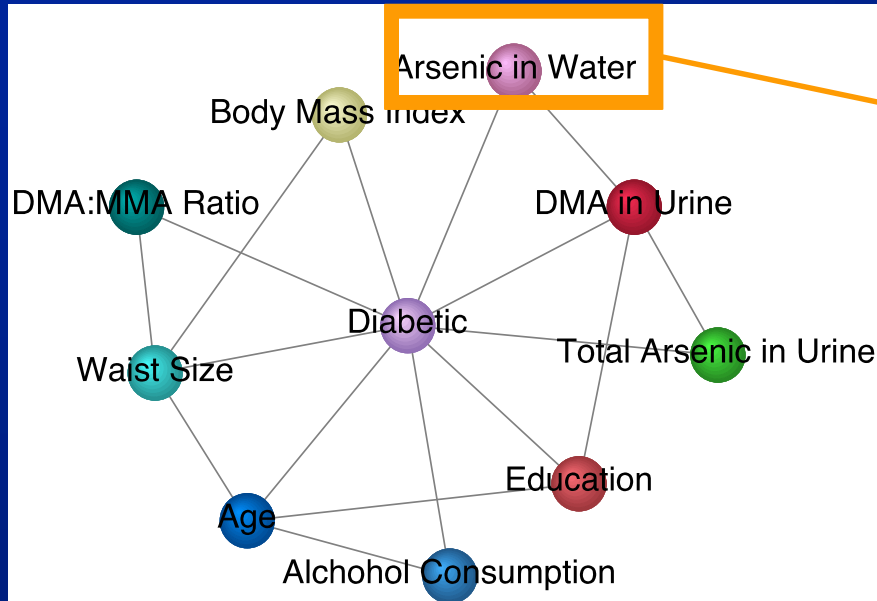




# Bayesian Network Can Estimate Diabetes Risk Reduction Benefits

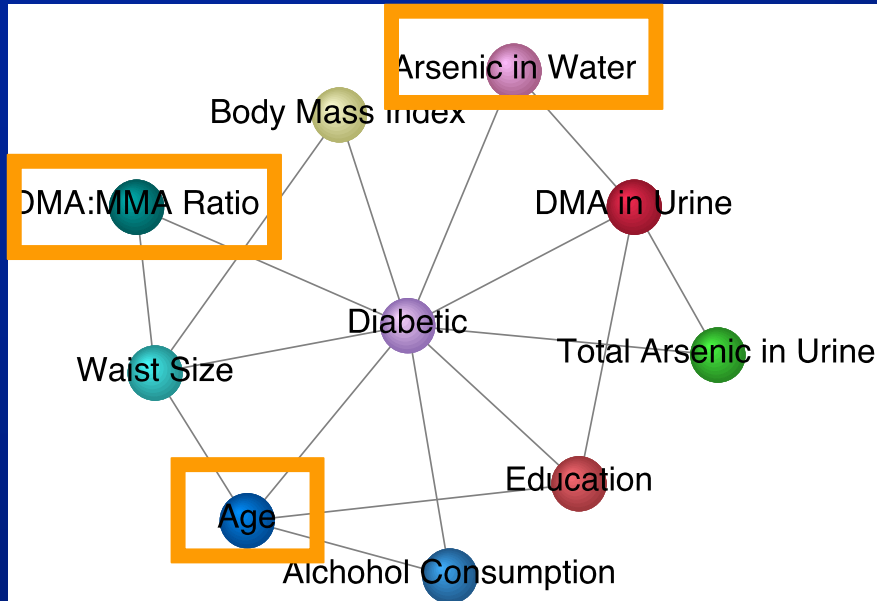


# Bayesian Network Can Estimate Diabetes Risk Reduction Benefits

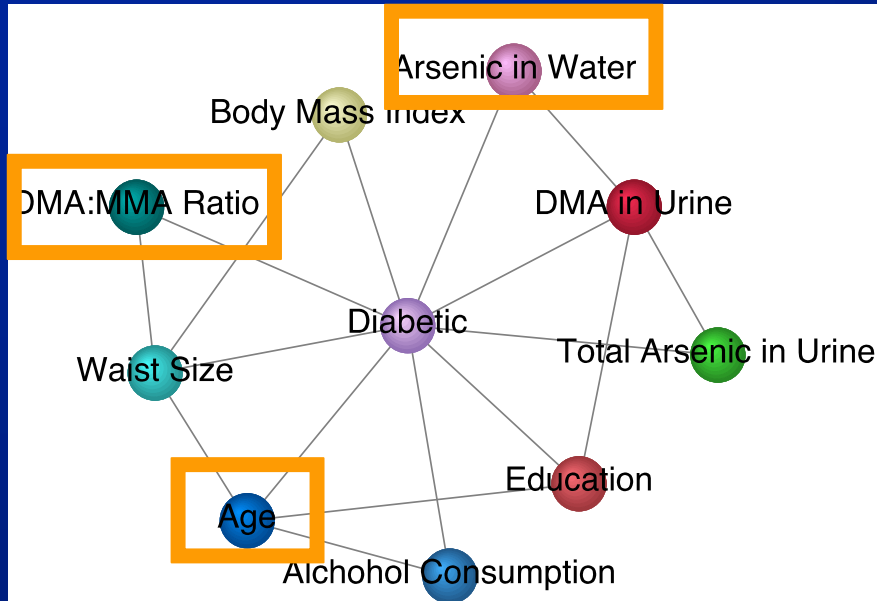


- $\Delta \text{Cases}$   
=  $BN_0 - BN_{\text{intervention}}$   
=  $170 - 160$   
**=10**

# Network Can Also Estimate Effects on Vulnerable Populations



# Network Can Also Estimate Effects on Vulnerable Populations



- Age > 55
- High arsenic methylation during metabolism
- $\Delta \text{Cases} = BN_{0, \text{vulnerable}} - BN_{\text{intervention, vulnerable}}$   
 $= 422 - 407$   
 $= 15$

# Future Platform for Dose-Response Assessment?

Bayesia Simulator Diabetes Risk from

**Age**

Mean  45.6914285714

☐ Observed

**Alcohol Consumption**

0  1

☐ Observed

**Arsenic in Water**

Mean  61.631860517

☐ Observed

**Diabetes Risk**

0 83.81%  
1 16.19%

**Body Mass Index**

Mean  28.984271362

☐ Observed

**DMA in Urine**

Mean  55.3494857868

☐ Observed

**DMA:MMA Ratio**

Mean  6.257525833

☐ Observed

**Education**

State

☐ Observed

**Total Arsenic in Urine**

Mean  74.5208286346

☐ Observed

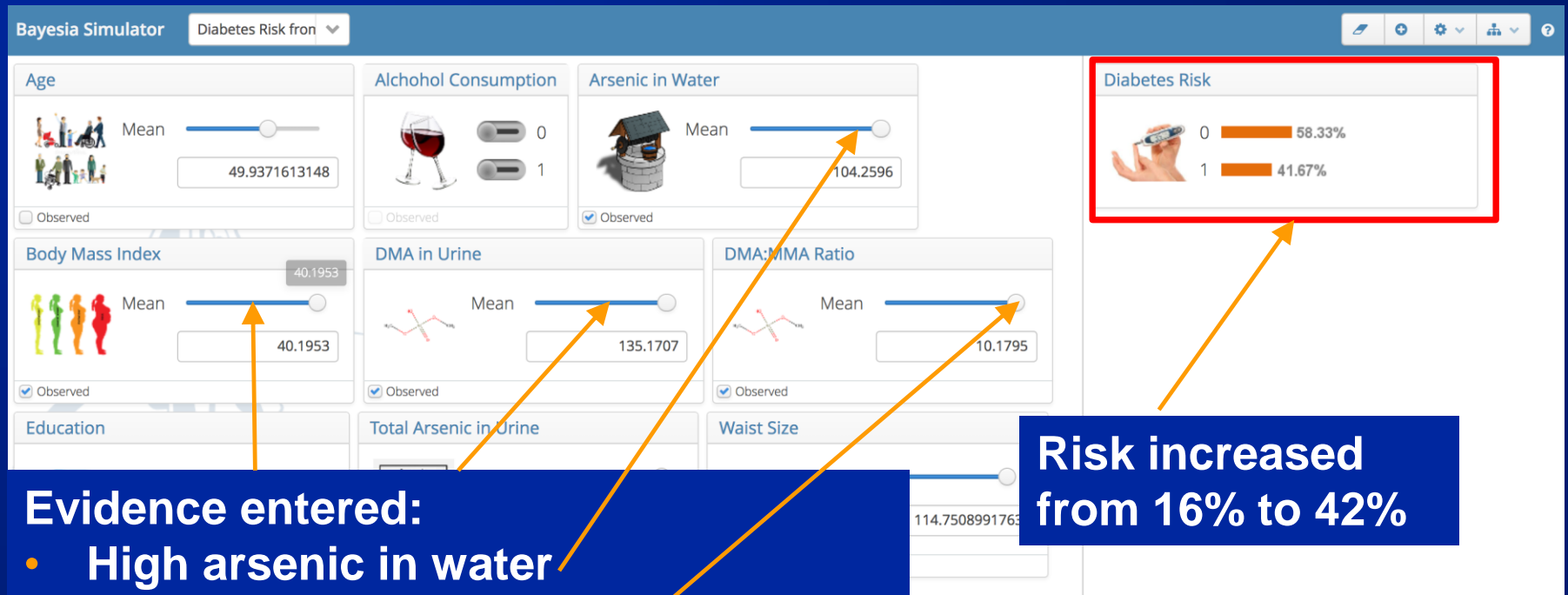
**Waist Size**

98.1472

Mean  98.1471850512

☐ Observed

# Interactive Updating with Evidence via a Web Platform




<https://simulator.bayesialab.com/> - !simulator/178925952810

# Differs from Current, Static Approach

## Noncancer Assessment

[Reference Dose for Oral Exposure \(RfD\) \(PDF\)](#) (29 pp, 186 K)

last updated: 09/01/1991

System	RfD (mg/kg-day)	Basis	PoD
 Cardiovascular, Dermal	$3 \times 10^{-4}$	Hyperpigmentation, keratosis and possible vascular complications	NOAEL : $8 \times 10^{-4}$ mg/kg-day

# Cancer Assessment

[Weight of Evidence for Cancer \(PDF\)](#) (29 pp, 186 K)

last updated: 06/01/1995

WOE Characterization	Framework for WOE Characterization
A (Human carcinogen)	Guidelines for Carcinogen Risk Assessment (US EPA, 1986)

## Basis:

- Based on sufficient evidence from human data. An increased lung cancer mortality was observed in multiple human populations exposed primarily through inhalation. Also, increased mortality from multiple internal organ cancers (liver, kidney, lung, and bladder) and an increased incidence of skin cancer were observed in populations consuming drinking water high in inorganic arsenic.
- This may be a synopsis of the full weight-of-evidence narrative.

---

[Quantitative Estimate of Carcinogenic Risk from Oral Exposure \(PDF\)](#) (29 pp, 186 K)

**Oral Slope Factor:** 1.5 per mg/kg-day

**Drinking Water Unit Risk:**  $5 \times 10^{-5}$  per  $\mu\text{g/L}$

**Extrapolation Method:** Time- and dose-related formulation of the multistage model

**Tumor site(s):** Dermal

**Tumor type(s):** Skin cancer (Tseng, 1977; Tseng et al., 1968; U.S. EPA, 1988)



# Summary

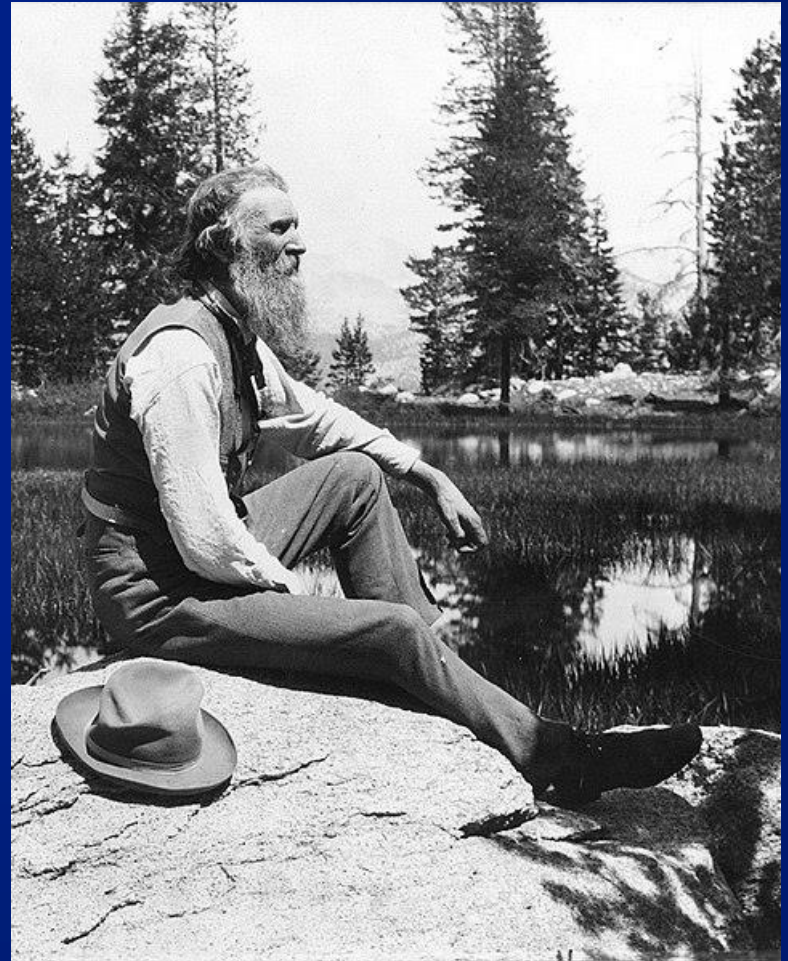
- **Current dose-response assessment methods:**
  - Are inconsistent for cancer, other illnesses
  - Do not capture non-linear relationships
  - Overlook complex interactions
    - E.g., genetics, environment
  - May have suboptimal discrimination capability
- **Bayesian networks could provide a new approach.**

**“When we try to pick out anything by itself, we find that it is bound fast by a thousand invisible cords that cannot be broken, to everything in the universe.”**

**John Muir, 1869**

**Naturalist**

**Sierra Club Founder**



# Acknowledgements

- **Joseph W. Zabinski (former PhD student, now at McKinsey & Co.)**
- **Chihuahua cohort data set:**
  - **University of North Carolina–Chapel Hill**
    - Michelle A. Mendez (UNC, Department of Nutrition)
    - Rebecca C. Fry (UNC, Department of Environmental Sciences & Engineering)
    - Miroslav Stýblo (UNC, Department of Nutrition)
    - Maria Ishida
    - Daniela Gutiérrez-Torres
    - R. Jesse Saunders
    - Zuzana Drobná
    - John Buse
    - Dana Loomis
  - **Universidad Autonoma De Chihuahua**
    - Blanca Sánchez-Ramírez
  - **Colegio de Médicos Cirujanos y Homeópatas del Estado de Chihuahua**
    - Damián Viniegra Morales
    - Francisco A. Baeza Terrazas
  - **Facultad de Medicina, Universidad Juárez del Estado de Durango**
    - Gonzalo García-Vargas
  - **Centro de Investigación y de Estudios Avanzados del Instituto Politécnico Nacional**
    - Luz Del Razo