



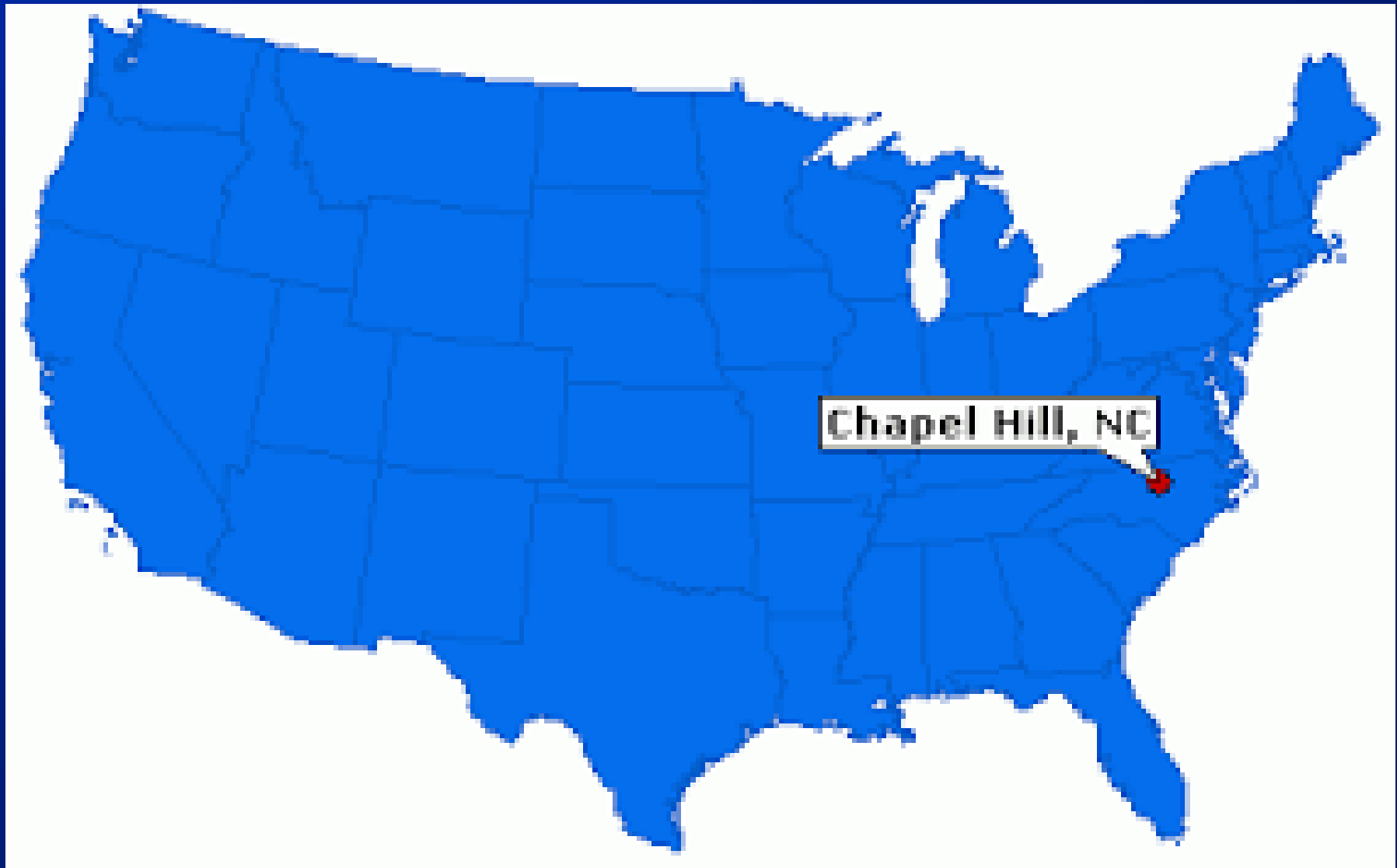
UNC
GILLINGS SCHOOL OF
GLOBAL PUBLIC HEALTH

Predicting Health Risks of Arsenic in Drinking Water

Bayesian Network Models

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

September 29, 2017



University of North Carolina, Chapel Hill, USA



- **Oldest public university in the United States**
 - 1789
 - 29,000 students, 3,850 faculty
- **Fourth school of public health in the United States**
 - 1940
 - 1,500 students, 250 faculty

Jacqueline MacDonald Gibson

- **Research focus: improved methods for environmental policymaking**
- **Academic background:**
 - **Ph.D., Civil and Environmental Engineering, Carnegie Mellon University, USA**
 - **Ph.D., Engineering and Public Policy, Carnegie Mellon University**
 - **M.S., Civil and Environmental Engineering, University of Illinois at Urbana-Champaign**
 - **B.A., Mathematics, Bryn Mawr College**

My Introduction to Bayesian Networks

Carnegie Mellon, 2003

Required class for all PhD students in Engineering and Public Policy

...making Bayesian networks more accessible to the probabilistically unsophisticated.

Bayesian Networks without Tears

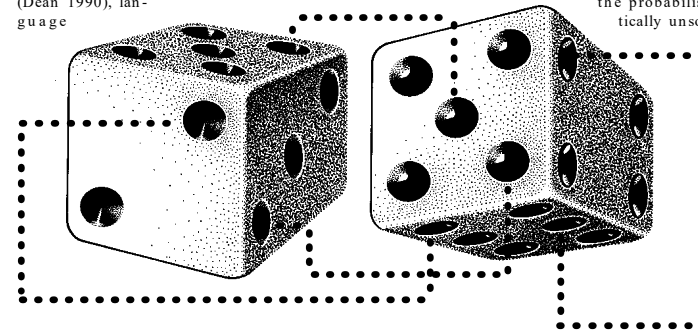
Eugene Charniak

Over the last few years, a method of reasoning using probabilities, variously called belief networks, Bayesian networks, knowledge maps, probabilistic causal networks, and so on, has become popular within the AI probability and uncertainty community. This method is best summarized in Judea Pearl's (1988) book, but the ideas are a product of many hands. I adopted Pearl's name, Bayesian networks, on the grounds that the name is completely neutral about the status of the networks (do they really represent beliefs, causality, or what?). Bayesian networks have been applied to problems in medical diagnosis (Heckerman 1990; Spiegelhalter, Franklin, and Bull 1989), map learning (Dean 1990), lan-

guage

I give an introduction to Bayesian networks for AI researchers with a limited grounding in probability theory. Over the last few years, this method of reasoning using probabilities has become popular within the AI probability and uncertainty community. Indeed, it is probably fair to say that Bayesian networks are to a large segment of the AI-uncertainty community what resolution theorem proving is to the AI-logic community. Nevertheless, despite what seems to be their obvious importance, the ideas and techniques have not spread much beyond the research community responsible for them. This is probably because the ideas and techniques are not that easy to understand. I hope to rectify this situation by making Bayesian networks more accessible to the probabilistically unsophisticated.

understanding (Charniak and Goldman 1989a, 1989b; Goldman 1990), vision (Levitt, Mullin, and Binford 1989), heuristic search (Hansson and Mayer 1989), and so on. It is probably fair to say that Bayesian networks are to a large segment of the AI-uncertainty community what resolution theorem proving is to the AI-logic community. Nevertheless, despite what seems to be their obvious importance, the ideas and techniques have not spread much beyond the research community responsible for them. This is probably because the ideas and techniques are not that easy to understand. I hope to rectify this situation by making Bayesian networks more accessible to the probabilistically unsophisticated.

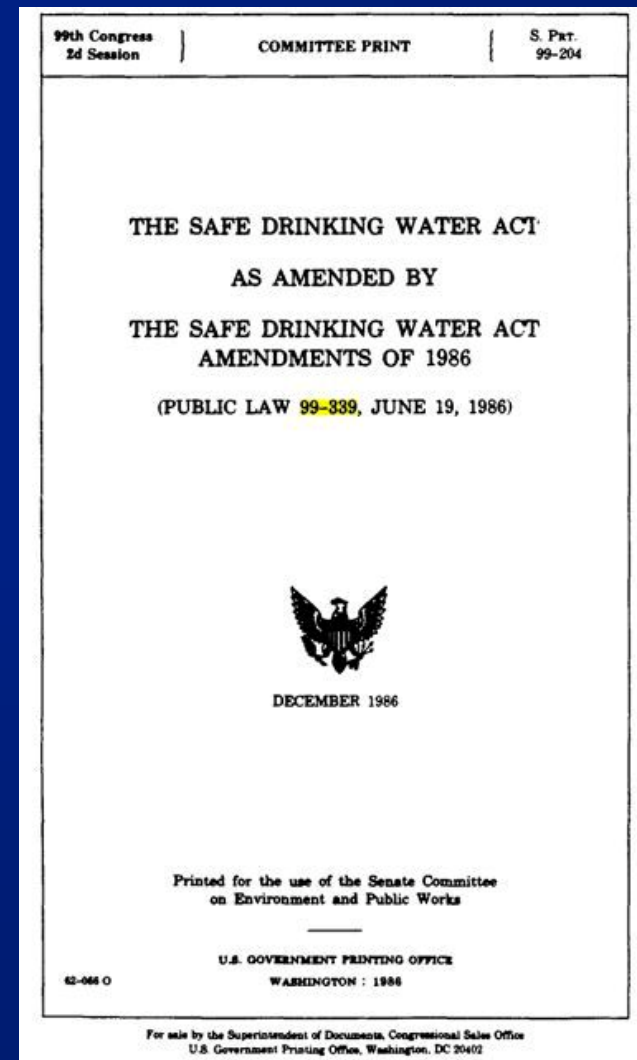


Outline

- **Introduction**
 - U.S. regulation of chemicals in drinking water
 - Arsenic in drinking water
 - Bayesian networks for improving arsenic risk assessment
- **Methods**
 - Low birthweight
 - Diabetes
- **Results: Bayesian network vs. traditional methods**
- **Discussion: Future vision for risk assessment of chemicals in water**

U.S. Safe Drinking Water Act Requires Risk Assessment of Chemicals

- “Maximum contaminant levels” are established via risk assessment
 - $< 1/10,000$ excess lifetime mortality risk
- Two key risk assessment steps:
 - Quantify chemical “dose”
 - Quantify lifetime illness risk associated with this dose



Environmental Protection Agency (EPA) Uses Two Methods to Calculate Risk

Cancer

$$P(\text{cancer}) = \alpha$$



“slope factor”

Other illnesses

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



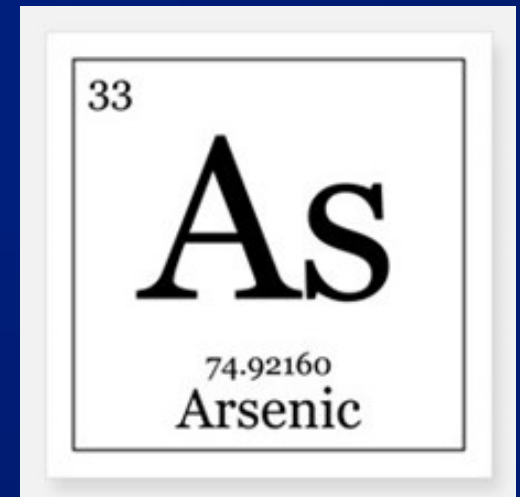
“reference dose”

EPA Approaches Have Many Limitations

- **Cancer and noncancer risk assessment methods differ**
 - Probability of illnesses other than cancer not quantified
- **Nonlinear relationships not captured**
- **Inter-individual variability (e.g., genetic differences) not captured**
- **Integration of evidence from multiple studies not possible**

Research Objectives

- Demonstrate Bayesian networks as alternative to current risk assessment approach
- Compare risk prediction capability to currently used methods
- Arsenic in water as case study



Arsenic Has Many Health Effects

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



Methods

Develop Bayesian Networks Using Data from Two Cohorts

Low birthweight

200 mothers and
infants



Gómez Palacio, Durango

Diabetes

1,050 adults



Chihuahua

Prior Study Associated Arsenic and Birthweight

Birthweight decreased as urinary mono-methylated arsenic (MMA) increased.

- Linear regression (Laine et al., 2015)



Maternal Metabolites

- Inorganic As
- **Monomethylated arsenic (MMA)**
- Dimethylated arsenic (DMA)



A Separate Study Associated Arsenic and Diabetes

Diabetes *decreased* as MMA increased but *increased* as DMA increased (Mendez et al., 2016).

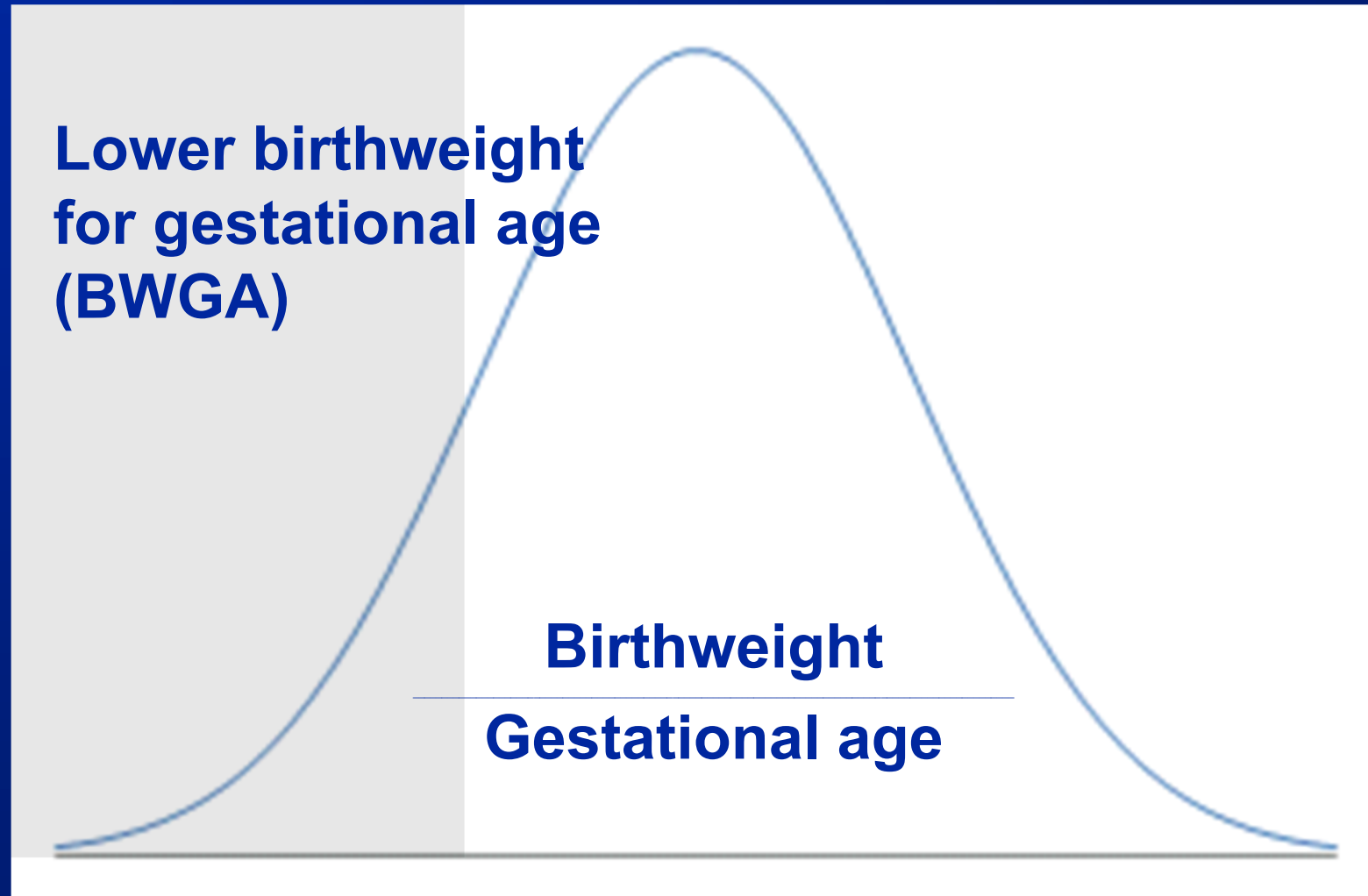


Metabolites

- Inorganic As
- Monomethylated arsenic (MMA)
- Dimethylated arsenic (DMA)



Analysis 1: Predict Lower Birthweight for Gestational Age



25th percentile

Analysis 2: Predict Diabetes Risk

- Diabetes defined according to World Health Organization guidelines:
 - Fasting plasma glucose ≥ 126 mg/dL
 - Two-hour plasma glucose ≥ 200 mg/dL
 - Self-reported diabetes diagnosis or medication use



Both Risks (Lower Birthweight, Diabetes) Assessed with BayesiaLab



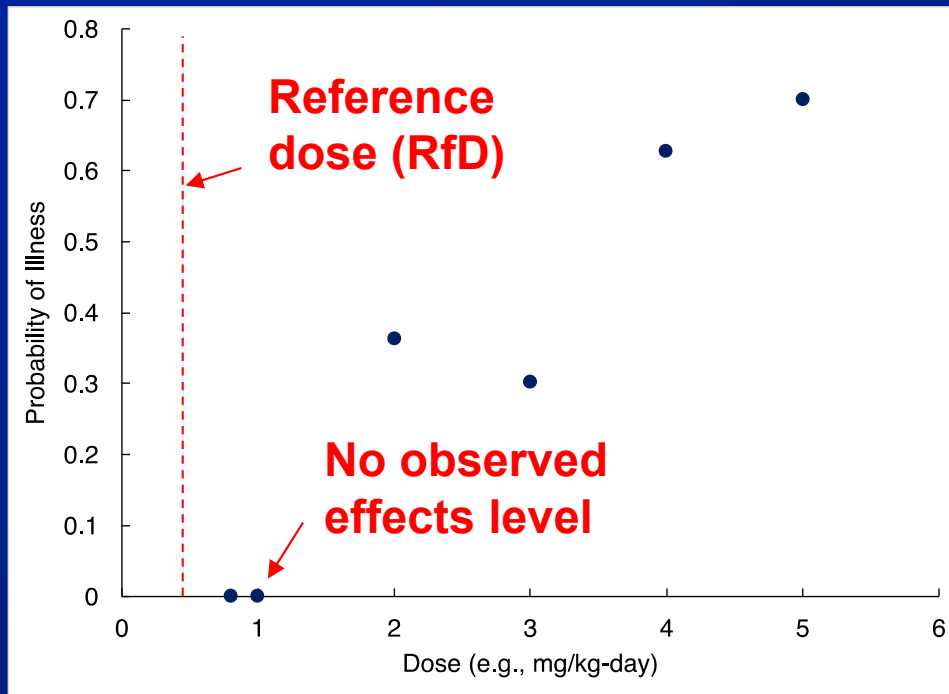
Lower Birthweight

- 11 predictor variables based on expert knowledge
- Continuous variables discretized into three states using R2-GenOpt
- Network structure developed through expert consultation

Diabetes

- 11 predictor variables
 - Discovered via unsupervised learning
- Continuous variables discretized into 5 states using R2-GenOpt
- Network structure and probability tables learned with augmented naïve Bayes algorithm

We Compared Bayes Net to EPA Method for Noncancer Risk Assessment

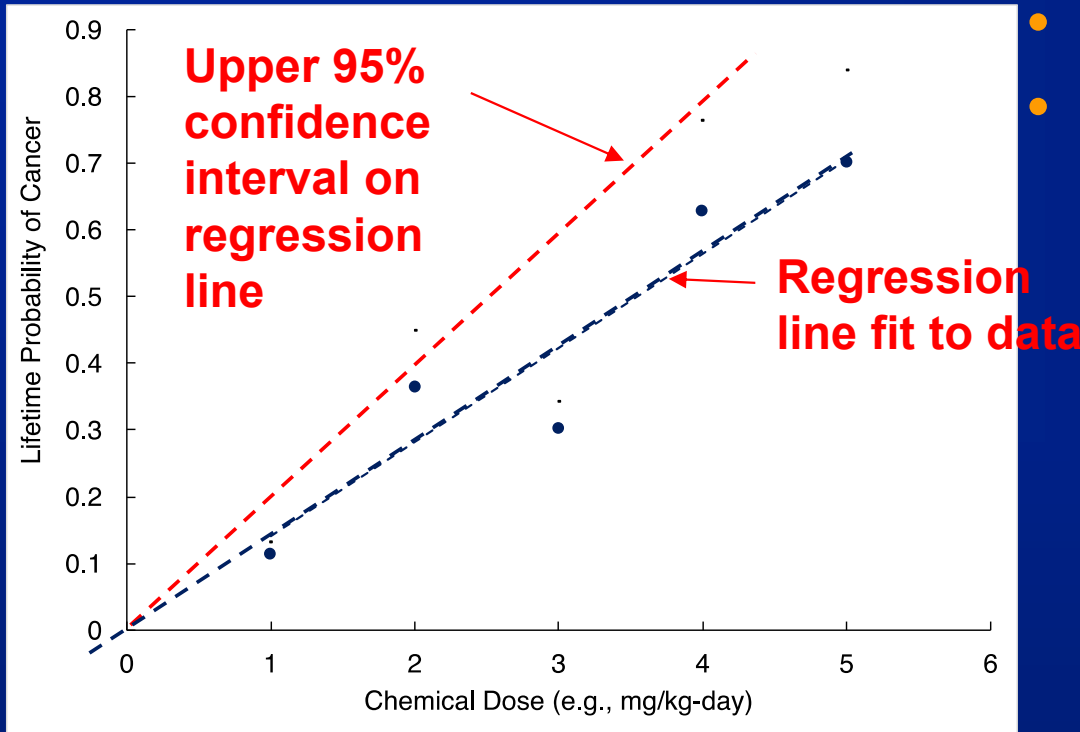


Compute reference dose:

- Divide no observed effects level by 3

$$P(illness) = \begin{cases} 0 & \text{if } \frac{Dose}{RfD} < 1 \\ 1 & \text{if } \frac{Dose}{RfD} \geq 1 \end{cases}$$

We Also Compared Bayes Net to EPA Cancer Risk Assessment Method

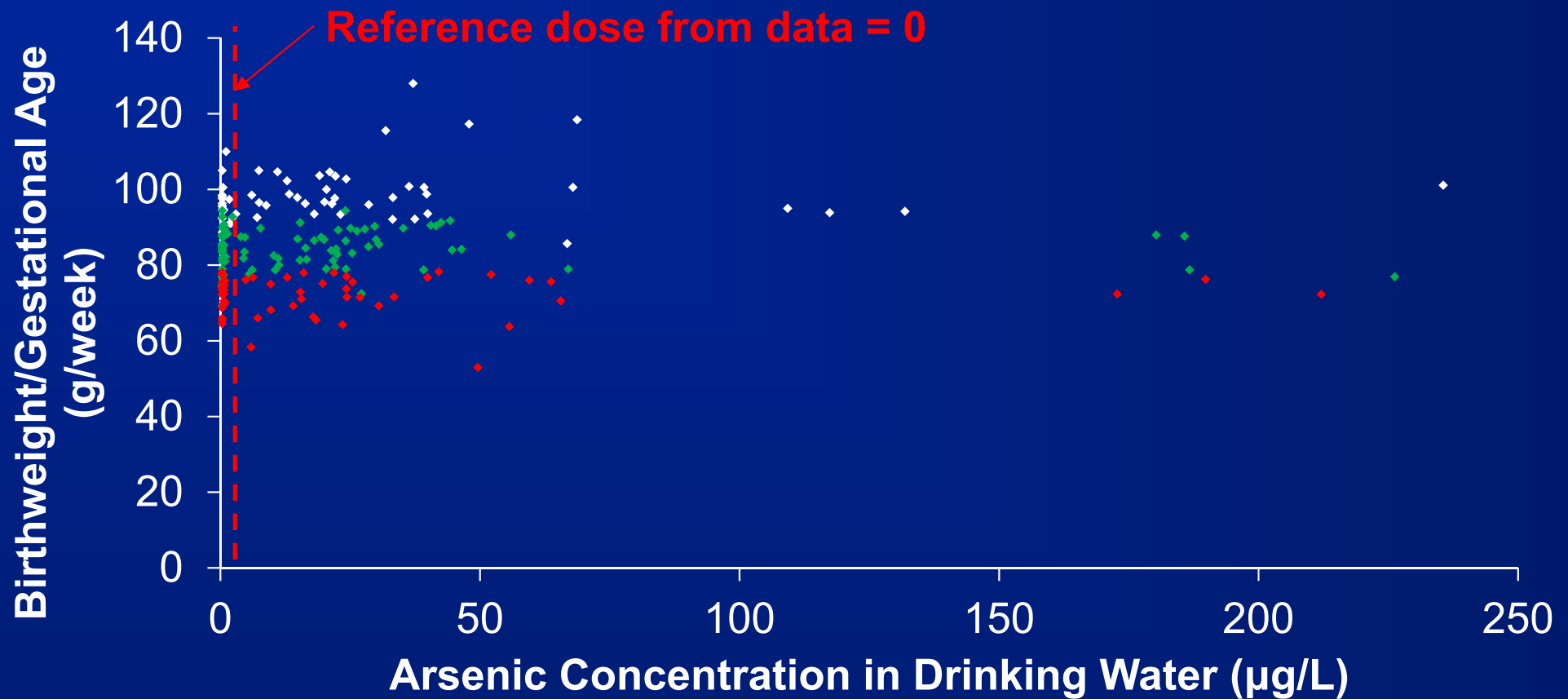


- Fit regression to the data.
- Compute upper 95% confidence interval on regression line slope.

$$P(cancer) = slope \times Dose$$

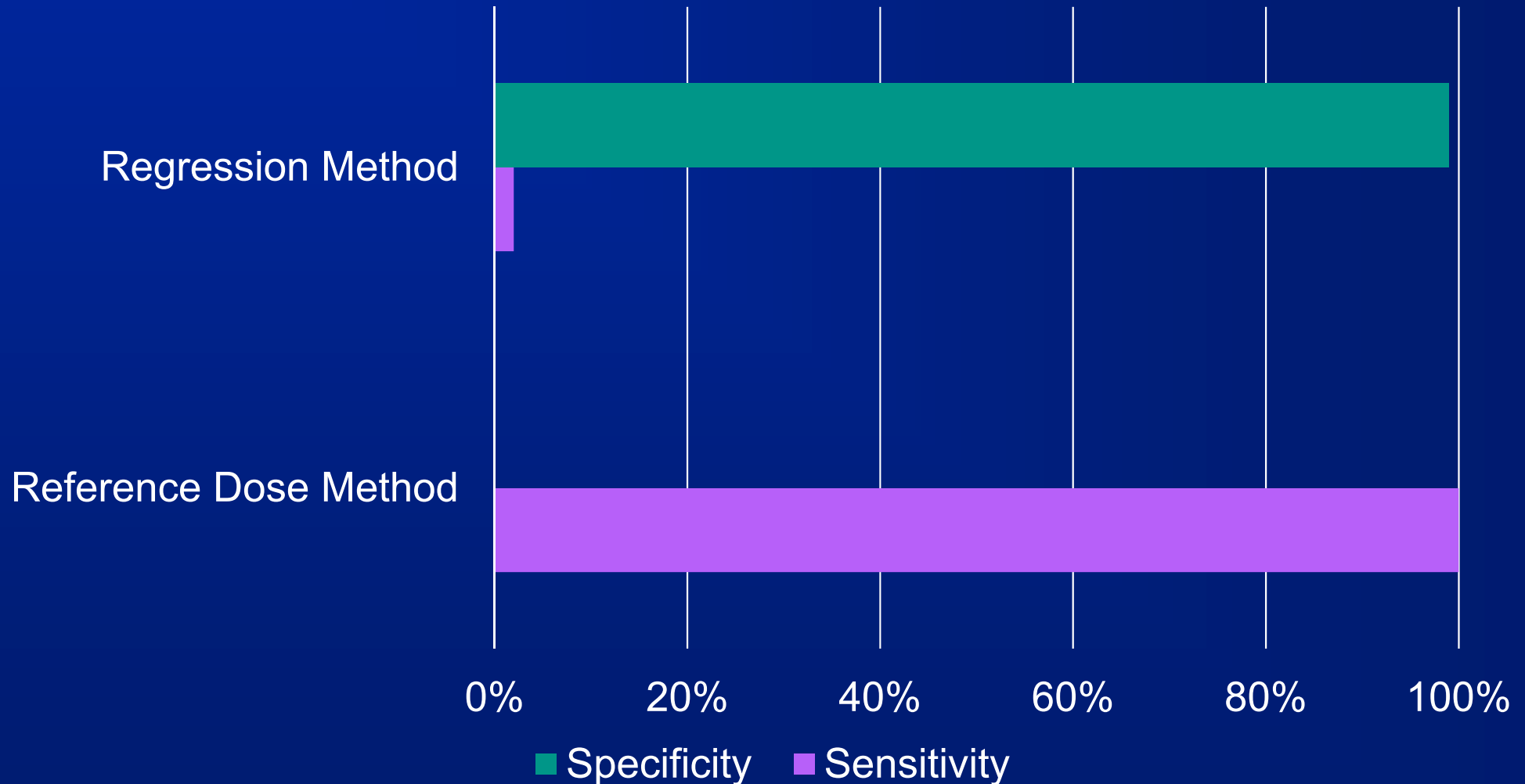
Results: Arsenic Association with Birthweight

No Clear “Safe Dose”—Lower Birthweights at Lowest Doses

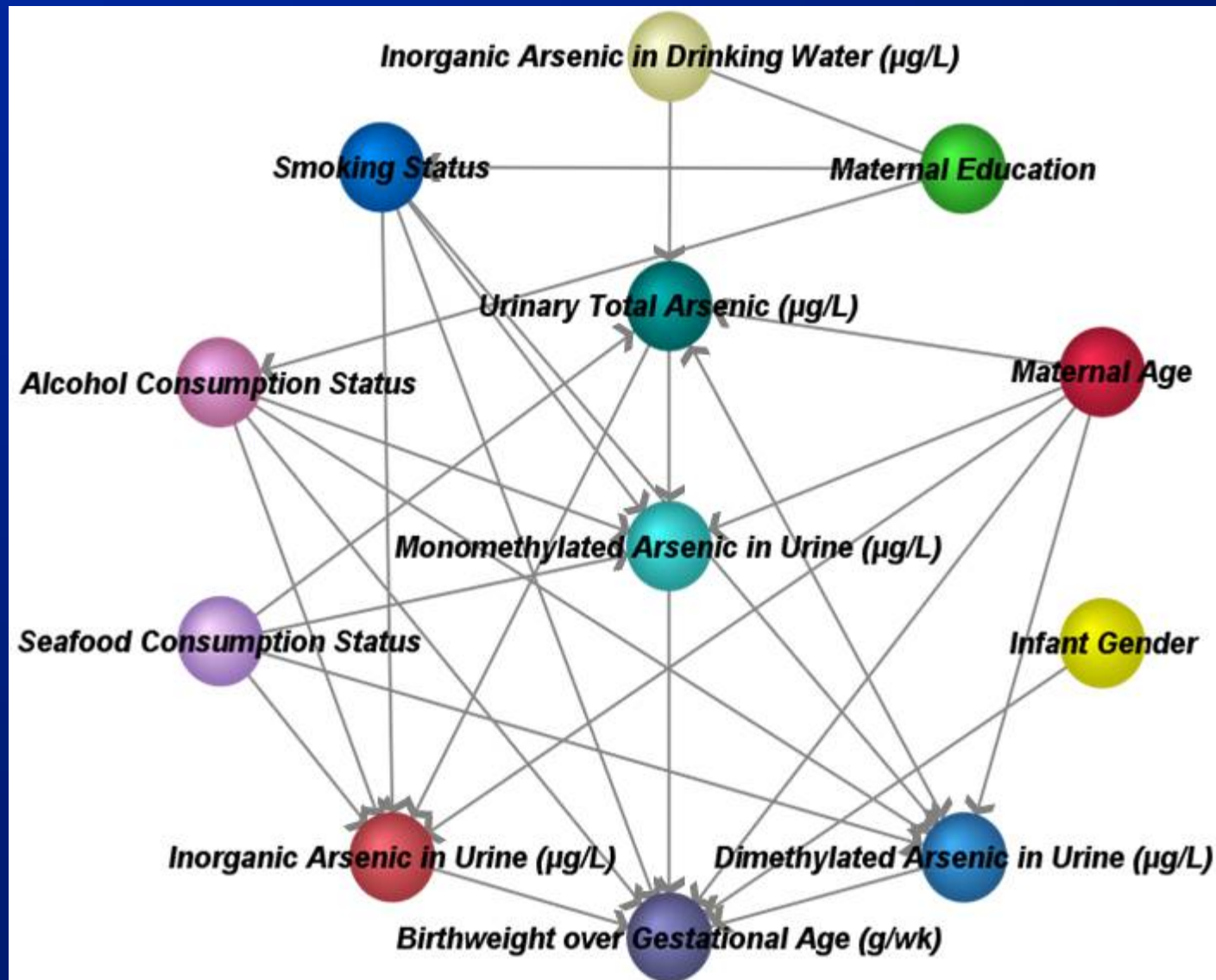


- large for gestational age (LGA)
- normal for gestational age
- small for gestational age (SGA)

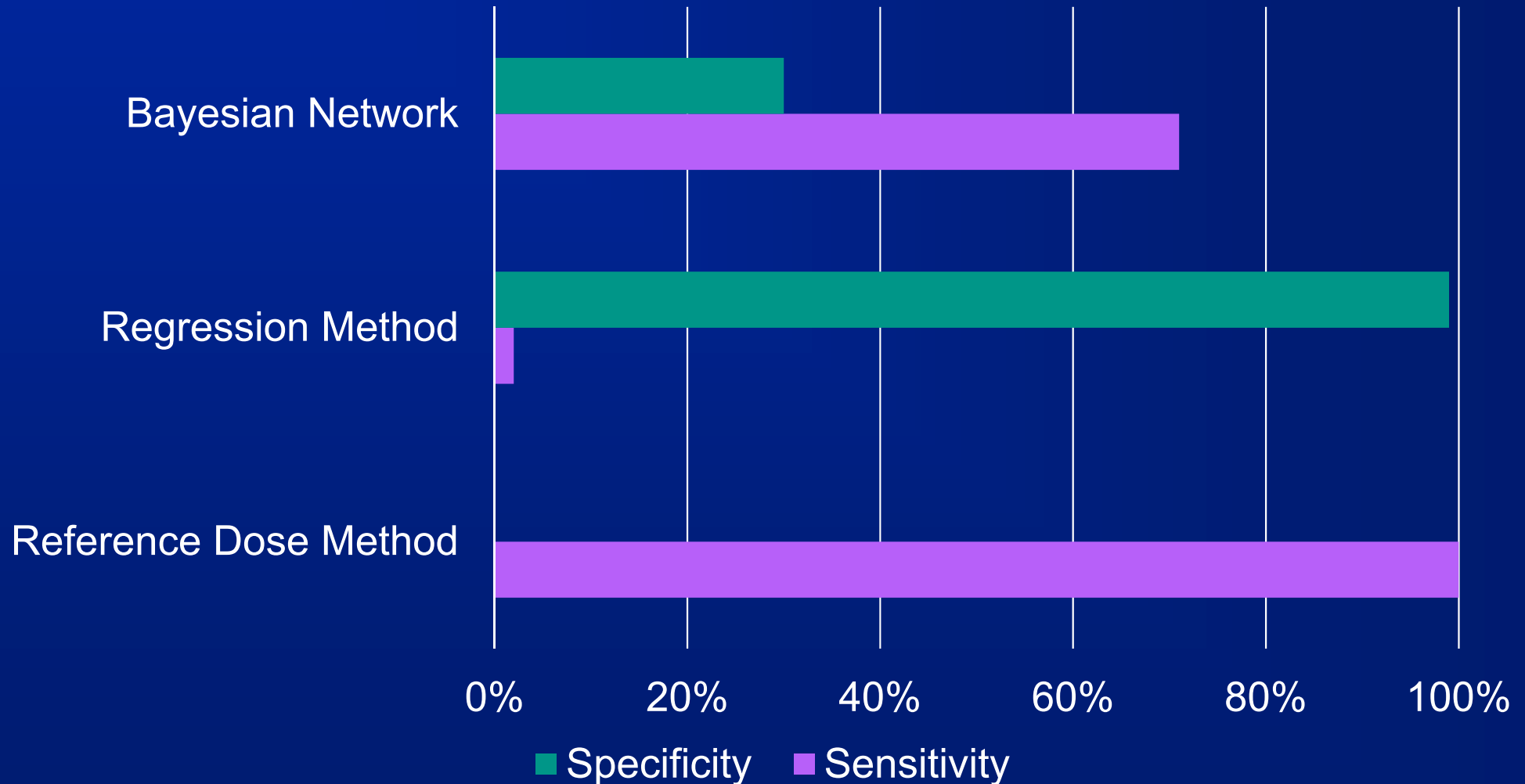
Existing Approaches Have Very Poor Discriminative Capability



Network Structure From Experts; Parameters from Data

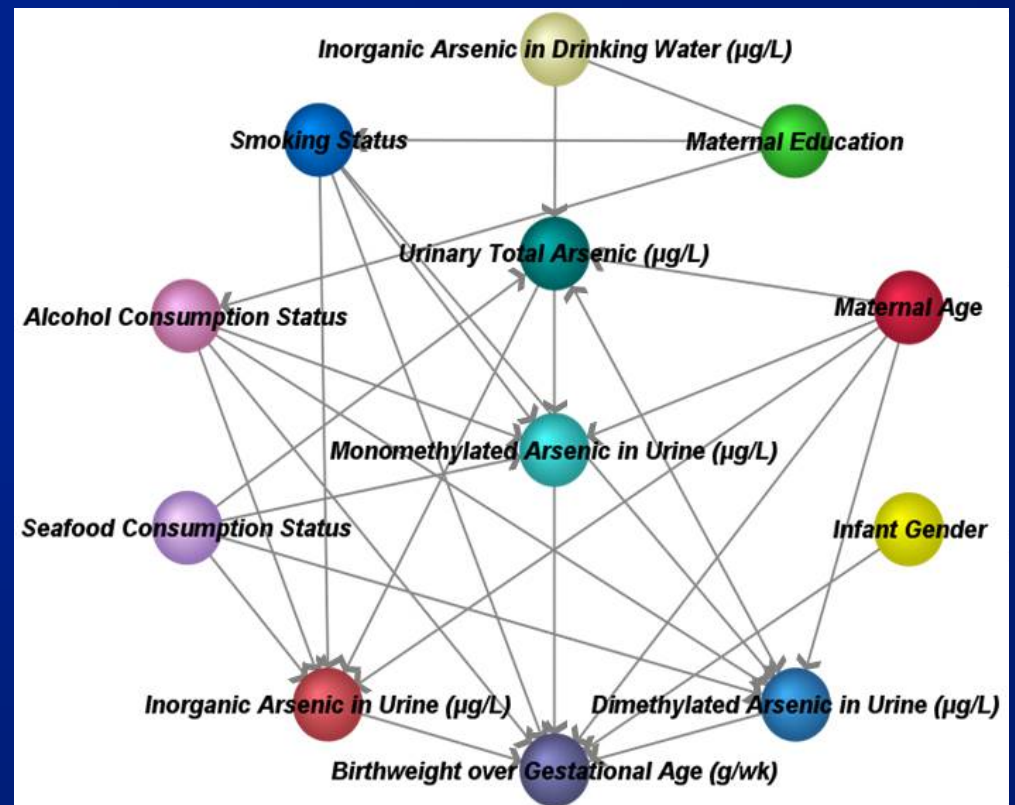


BayesiaLab Model Improves Discriminative Capability



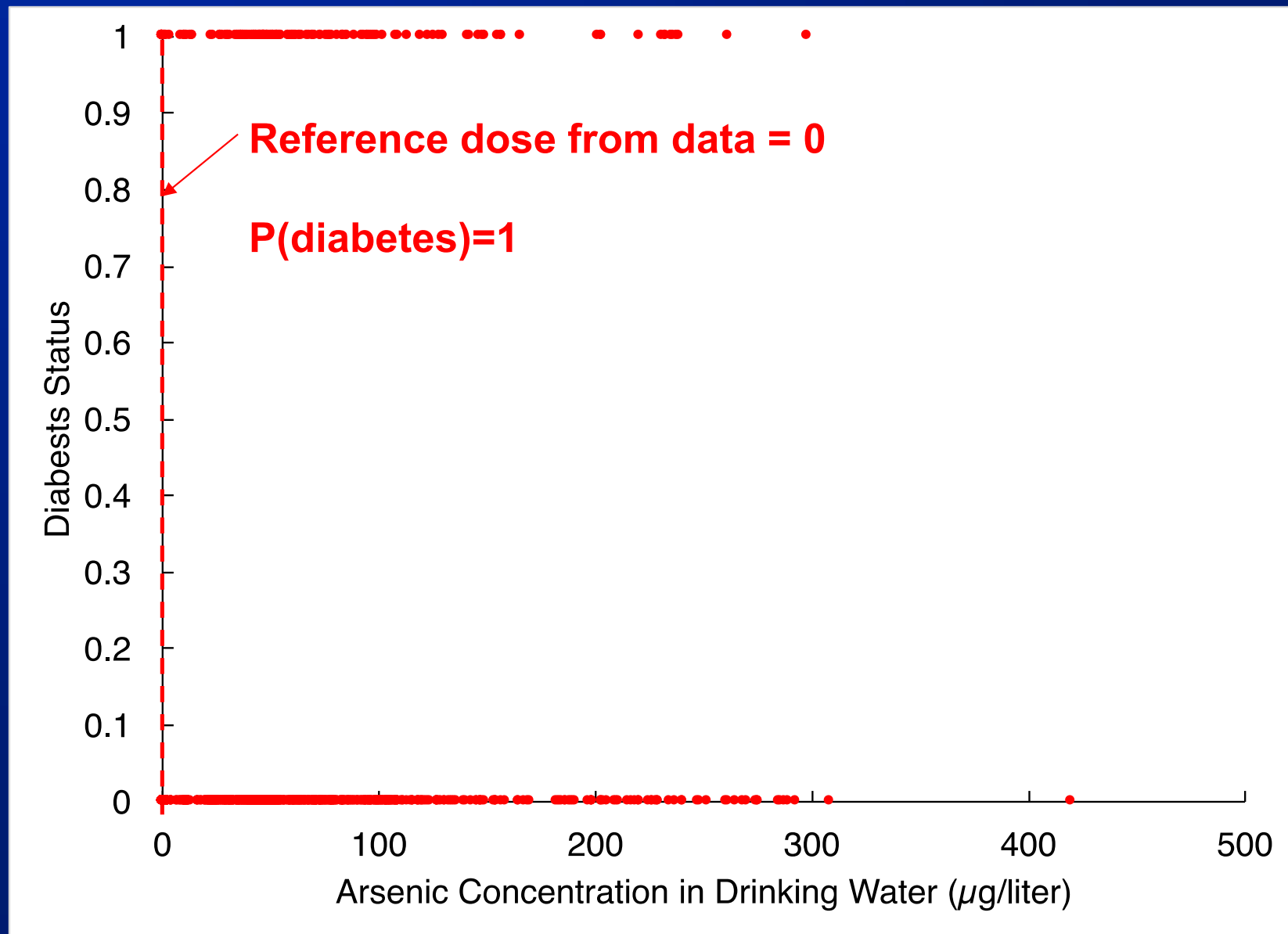
Future Steps

- Automated learning of network structure
 - Existing structure is based on expert beliefs

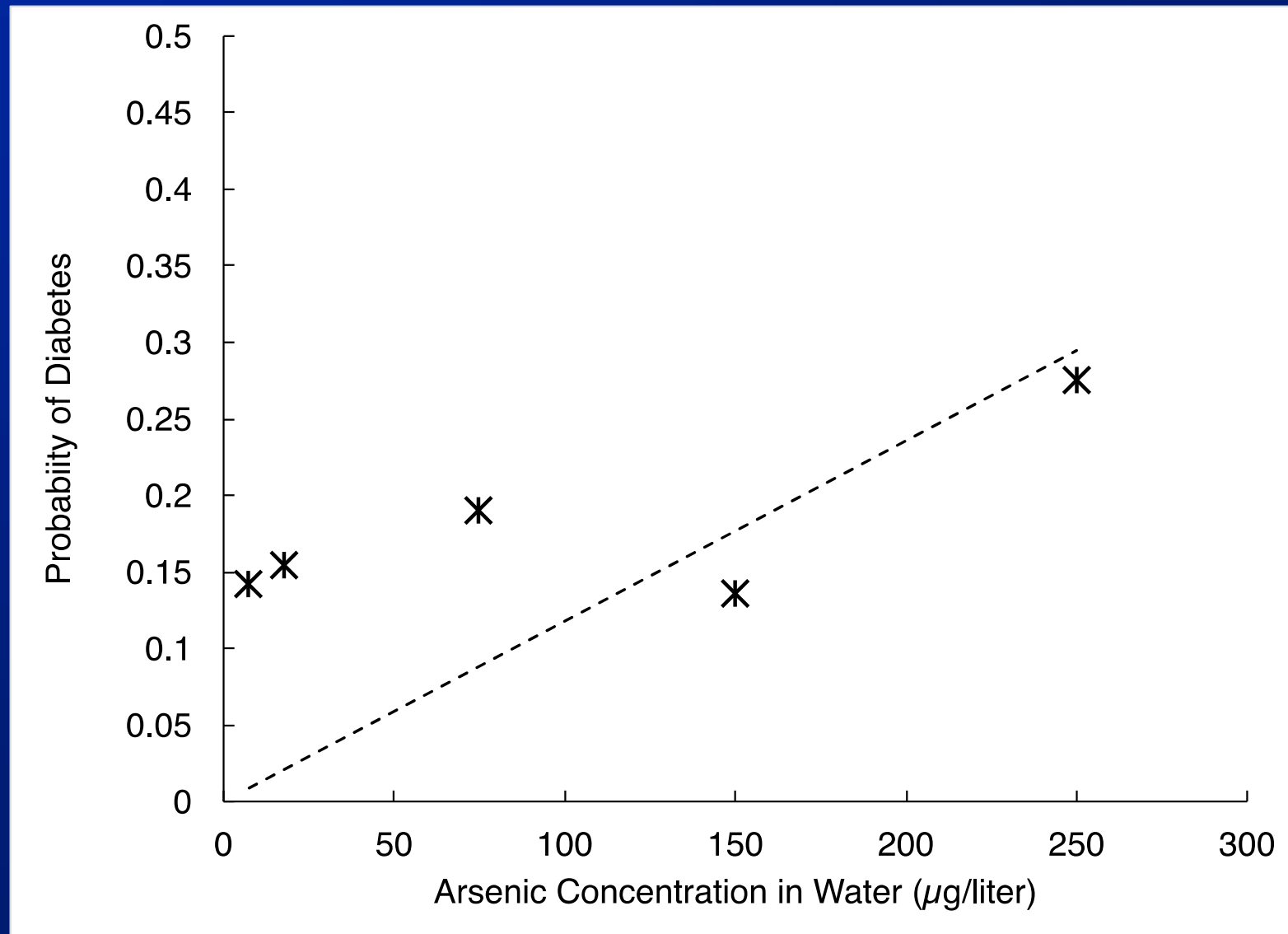


Results: Arsenic Association with Diabetes

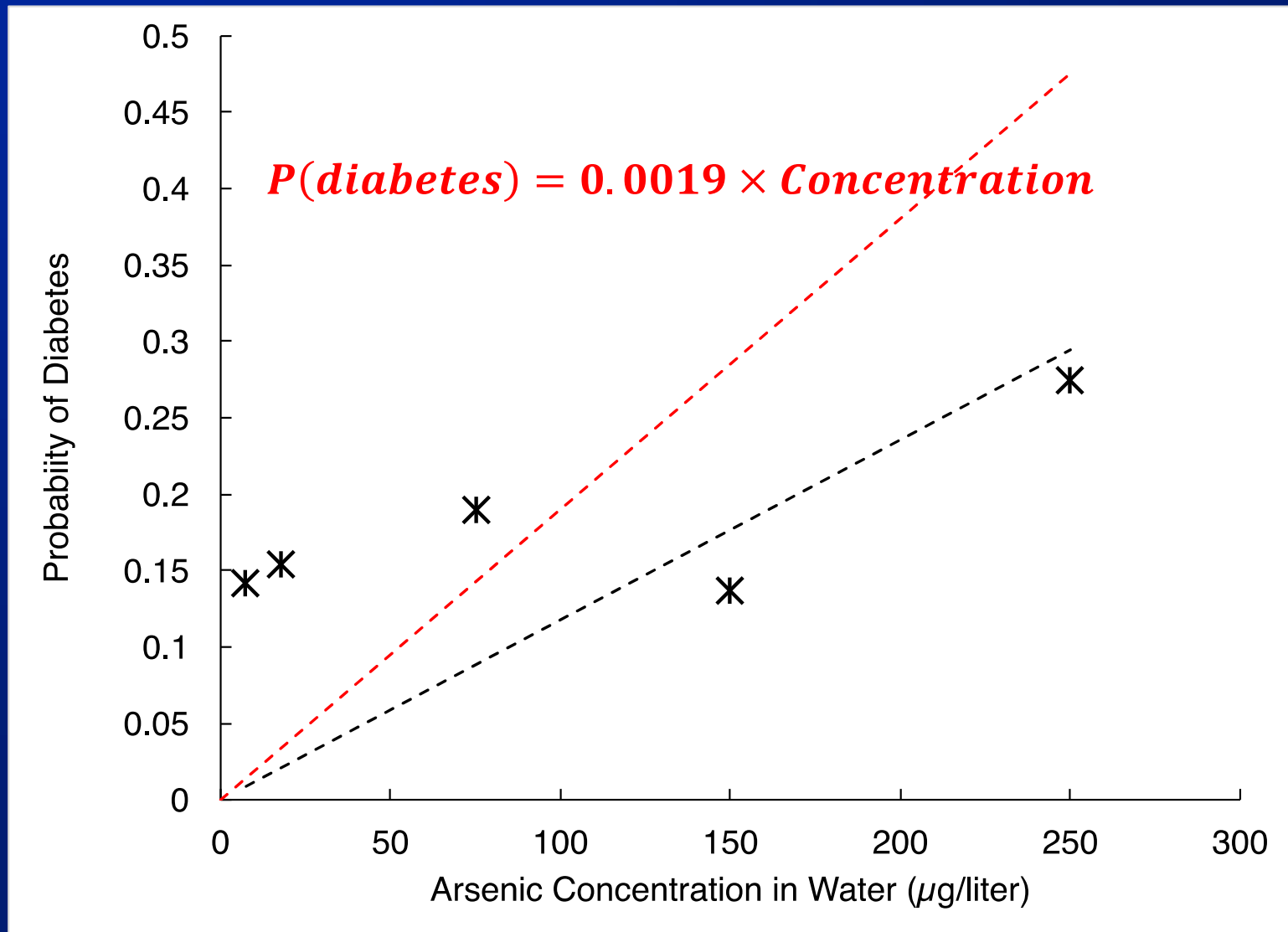
Reference Dose Based on Data Is Zero — “Everyone At Risk”



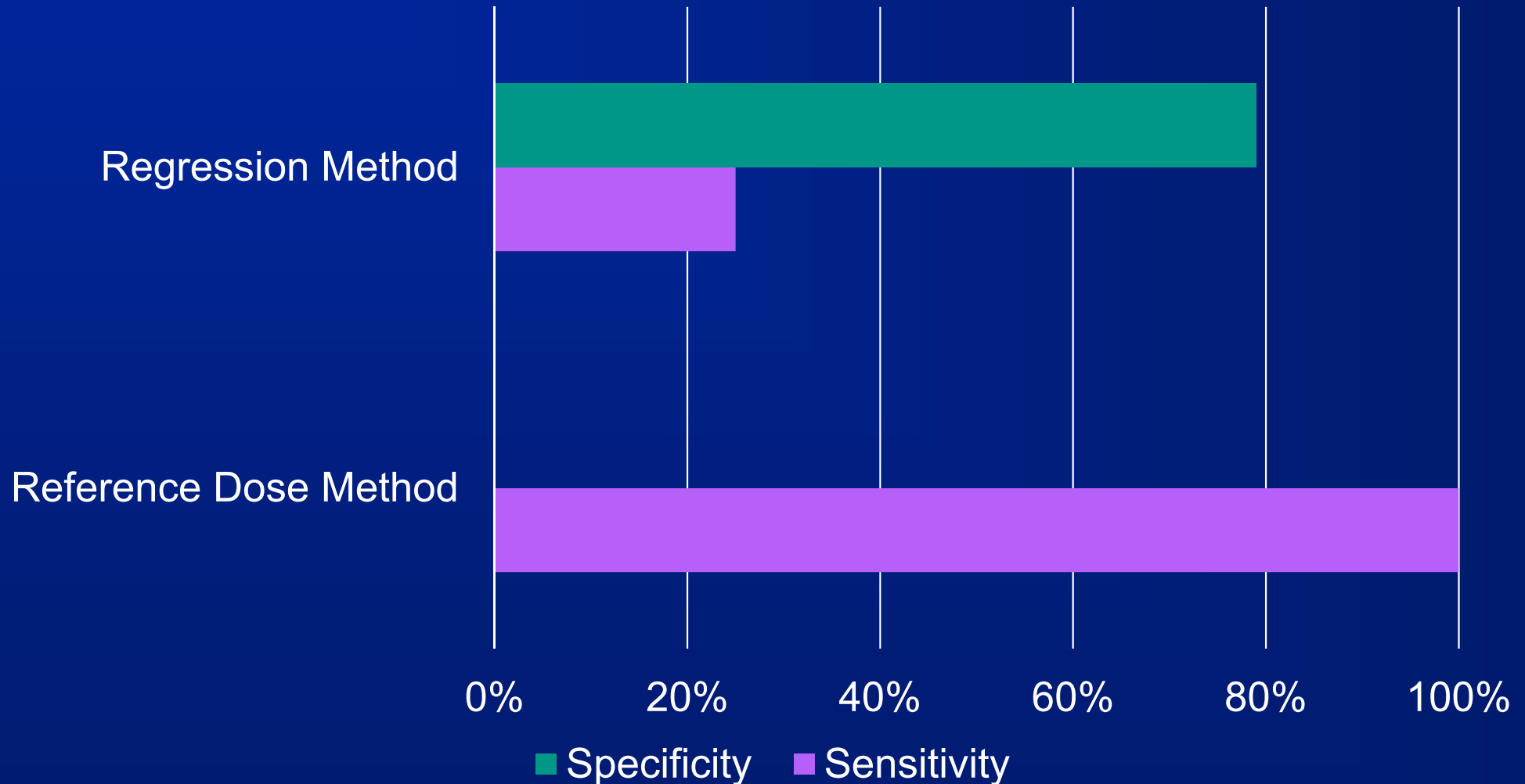
Group Cohort Into Dose Groups to Estimate “Slope Factor”



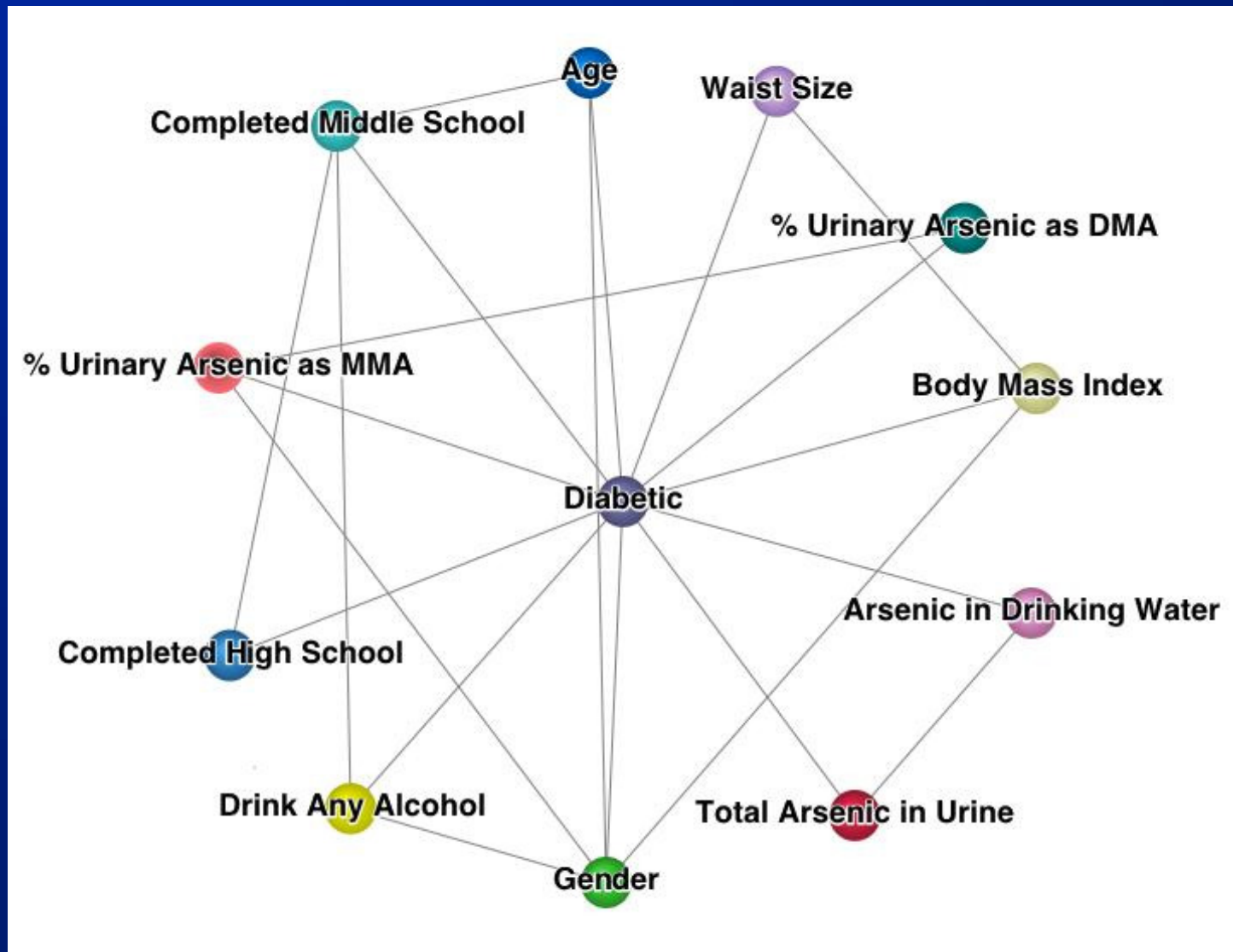
Group Cohort Into Dose Groups to Estimate “Slope Factor”



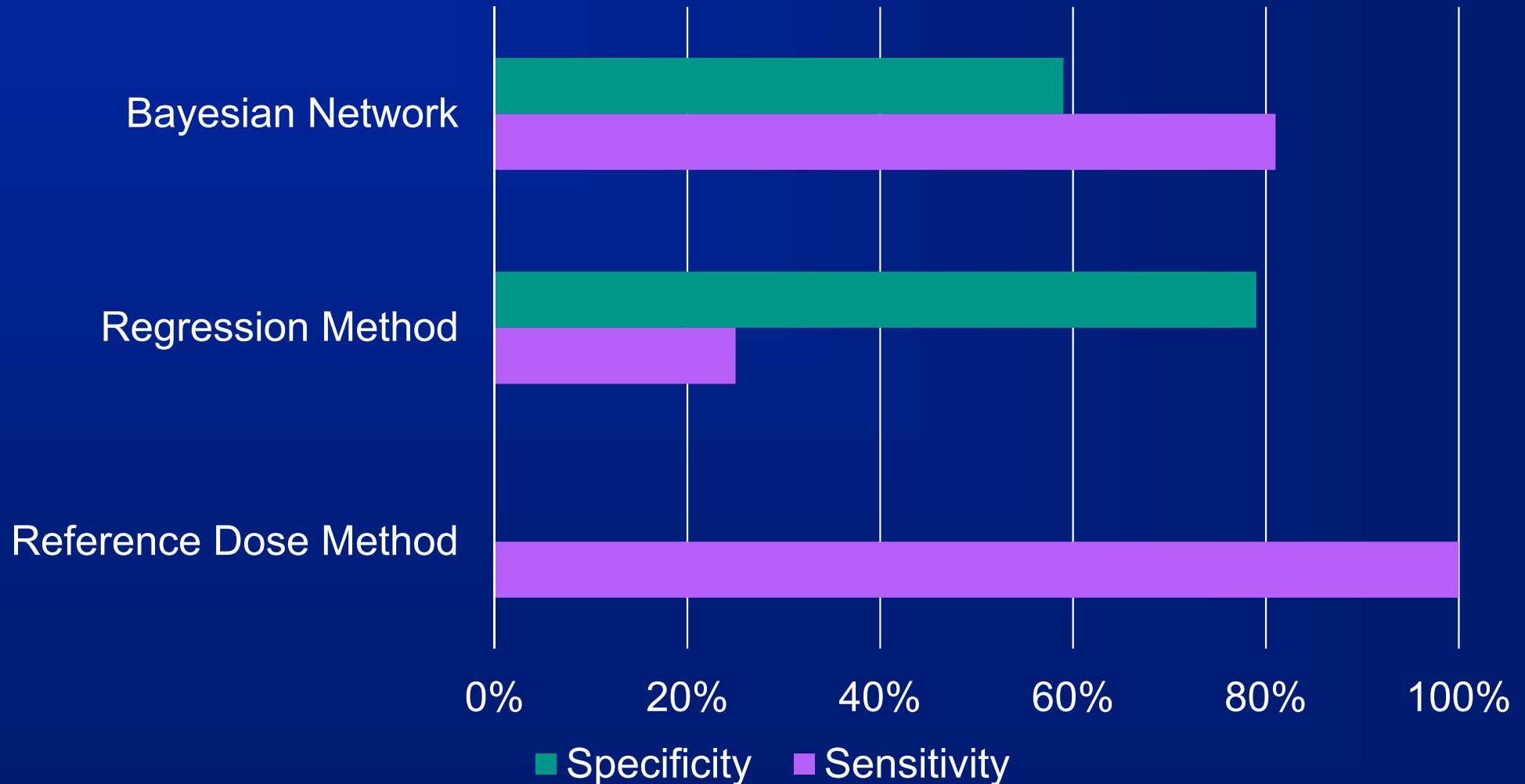
As for Birthweight, Existing Methods Have Poor Discrimination Ability



Machine-Learned Structure Consistent with Prior Knowledge

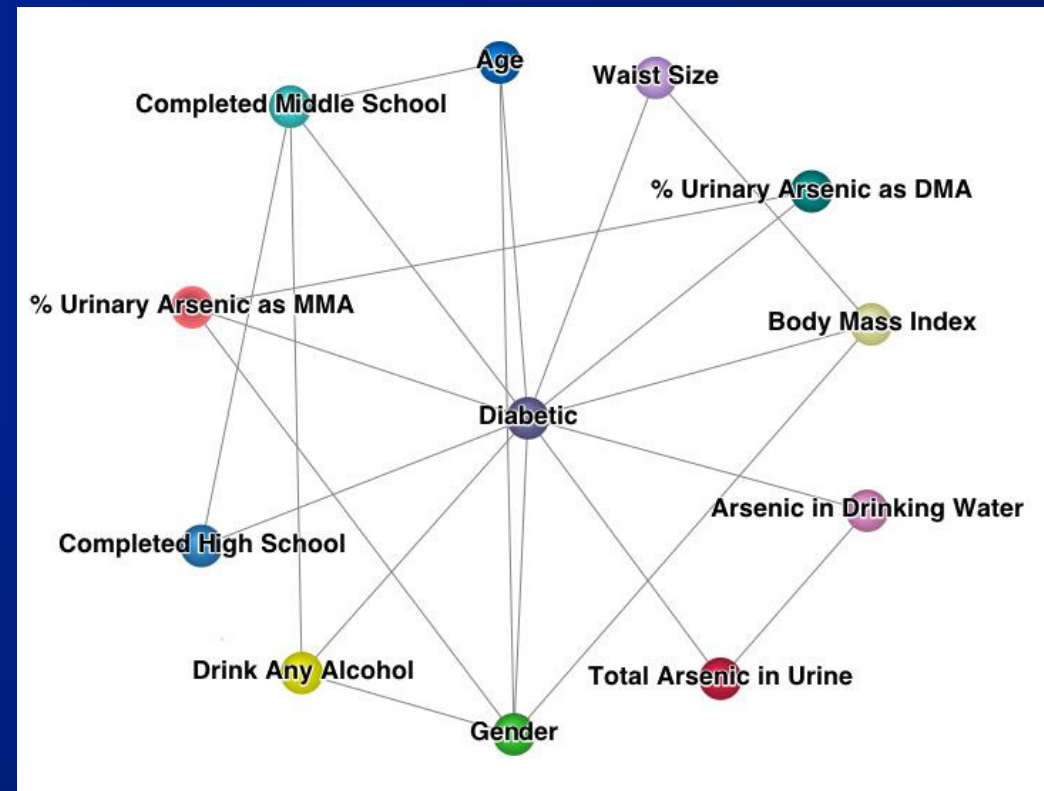


Bayesian Network Greatly Improves Discrimination Capability



Future Steps

- Optimize network structure.
 - Improved performance under cross-validation?



Conclusions

Summary

- **Current U.S. EPA methods for health risk assessment:**
 - Are inconsistent for cancer and noncancer illnesses
 - Have poor discrimination capability
 - Cannot be customized based on age, gender, genetics, etc.
- **Bayesian networks could provide a new approach.**

How Risk Assessors Quantify Risks in Current Practice

Step 1: Look up chemical information on EPA Integrated Risk Information System (IRIS) web site

The screenshot shows the EPA Integrated Risk Information System (IRIS) website. At the top is the EPA logo and navigation links: Environmental Topics, Laws & Regulations, and About EPA. A search bar is located on the right. The main heading is "Integrated Risk Information System". Below this, there is a section titled "IRIS Assessments in Review" with a list of two items: "Ethyl Tertiary Butyl Ether (ETBE) (External Review Draft)" and "Tert-butyl Alcohol (tert-butanol) (External Review Draft)". To the right of this section is a "Staying Connected" box with links for "How IRIS connects with you" and "How you can connect with IRIS". Below the "Staying Connected" box is a "Get email alerts" section with a text input field and a "sign up" button. The background of the "IRIS Assessments in Review" section shows a group of people sitting at tables in a meeting room.

EPA United States Environmental Protection Agency

Environmental Topics Laws & Regulations About EPA Search EPA.gov

Integrated Risk Information System

CONTACT US SHARE

IRIS Assessments in Review

- [Ethyl Tertiary Butyl Ether \(ETBE\) \(External Review Draft\)](#)
- [Tert-butyl Alcohol \(tert-butanol\) \(External Review Draft\)](#)

Staying Connected

- [How IRIS connects with you](#)
- [How you can connect with IRIS](#)


Get email alerts

sign up

Step 2: Look up Reference Dose, and Compute Noncancer “Risk”

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

If hazard quotient > 1,
then assume all are at
risk of noncancer
effects.

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 x 10 ⁻⁴	Hyperpigmentation, keratosis and possible vascular complications	NOAEL : 8 x 10 ⁻⁴ mg/kg-day
Reference Concentration for Inhalation Exposure (RfC) (PDF) (29 pp, 186 K) Not assessed under the IRIS Program.			

Step 3: Look up Slope Factor, and Compute Cancer Risk

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: 5×10^{-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)

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

New Vision: Risk Assessment via Bayesian Network Web Simulator

Bayesia Simulator Diabetes Risk from ▼ i ✎ +


% Urinary Arsenic as DMA



Mean

☐ Observed

Arsenic in Drinking Water



Mean

☐ Observed

Completed Middle School


MIDDLE

SCHOOL

☐ 0
☐ 1

☐ Observed


% Urinary Arsenic as MMA



Mean

☐ Observed


Body Mass Index



Mean

☐ Observed


Drink Any Alcohol



☐ 0
☐ 1

☐ Observed

Diabetic



0	<div><div></div></div>	83.61%
1	<div><div></div></div>	16.39%

Risk Assessor Could Enter Multiple Characteristics for Customized Estimate

Bayesia Simulator


Diabetes Risk from ▼

i

✎

+


% Urinary Arsenic as DMA



Mean

☒ Observed


% Urinary Arsenic as MMA



Mean

☐ Observed


Diabetic



0 50.13%

1 49.87%


Arsenic in Drinking Water



Mean

☒ Observed


Body Mass Index



Mean

☒ Observed

Completed Middle School




☐ 0

☐ 1

☒ Observed

Drink Any Alcohol



☐ 0

☐ 1

☐ Observed

Receiver-Operating Characteristic Curve Shows Discrimination Strength

