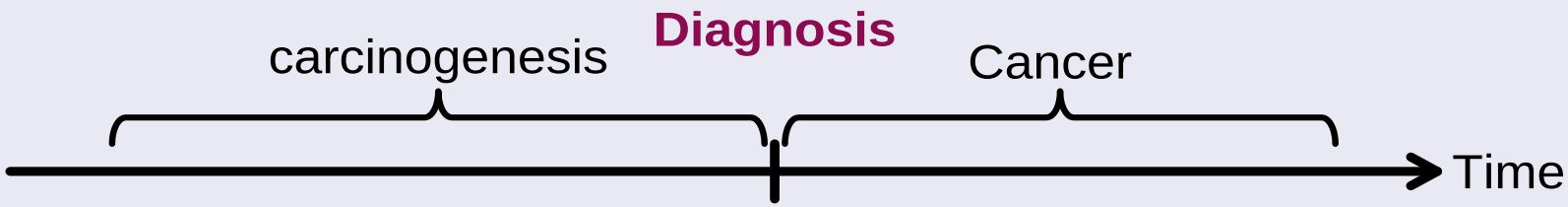
A statistical approach for the multi-stage model of carcinogenesis

Sandra Plancade, University of Tromso, Norway.
Gregory Nuel, University Paris-Descartes
Eiliv Lund, University of Tromso.

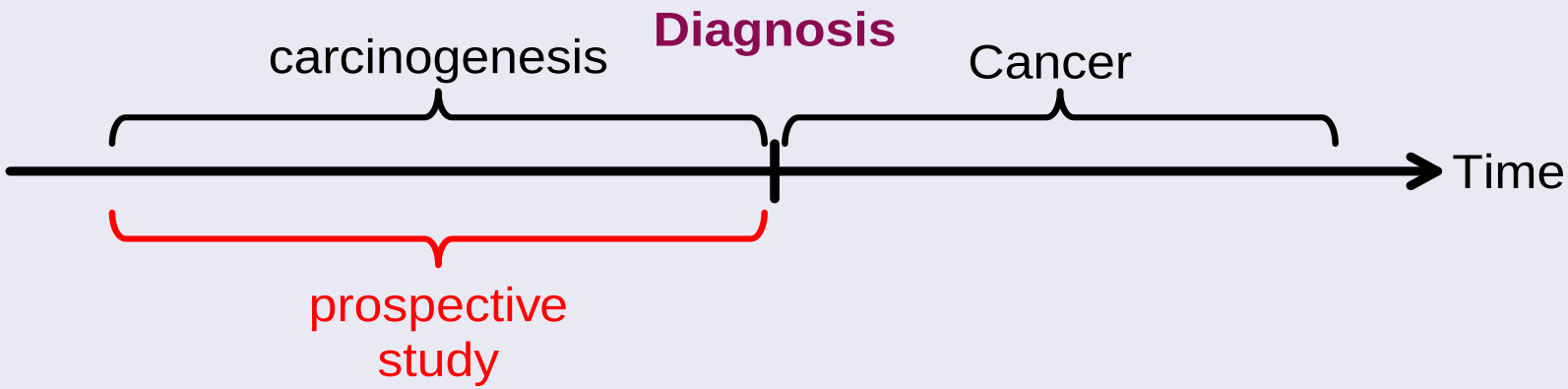
28 February 2012

Epidemiology and Carcinogenesis



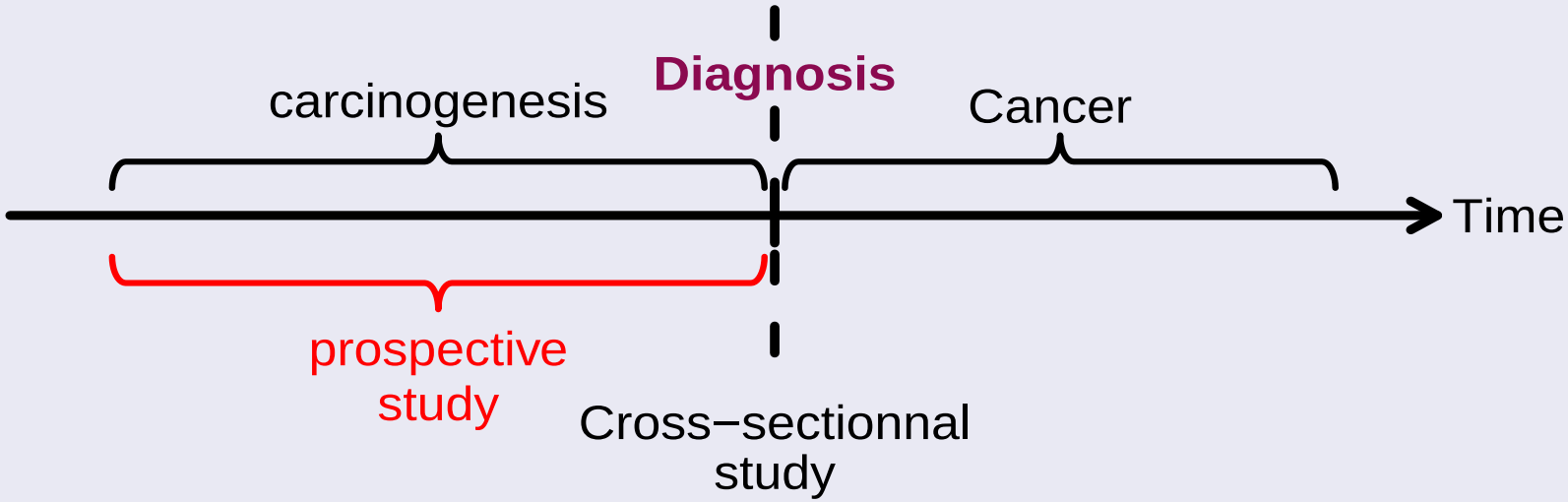
- Carcinogenesis: prior to diagnosis

Epidemiology and Carcinogenesis



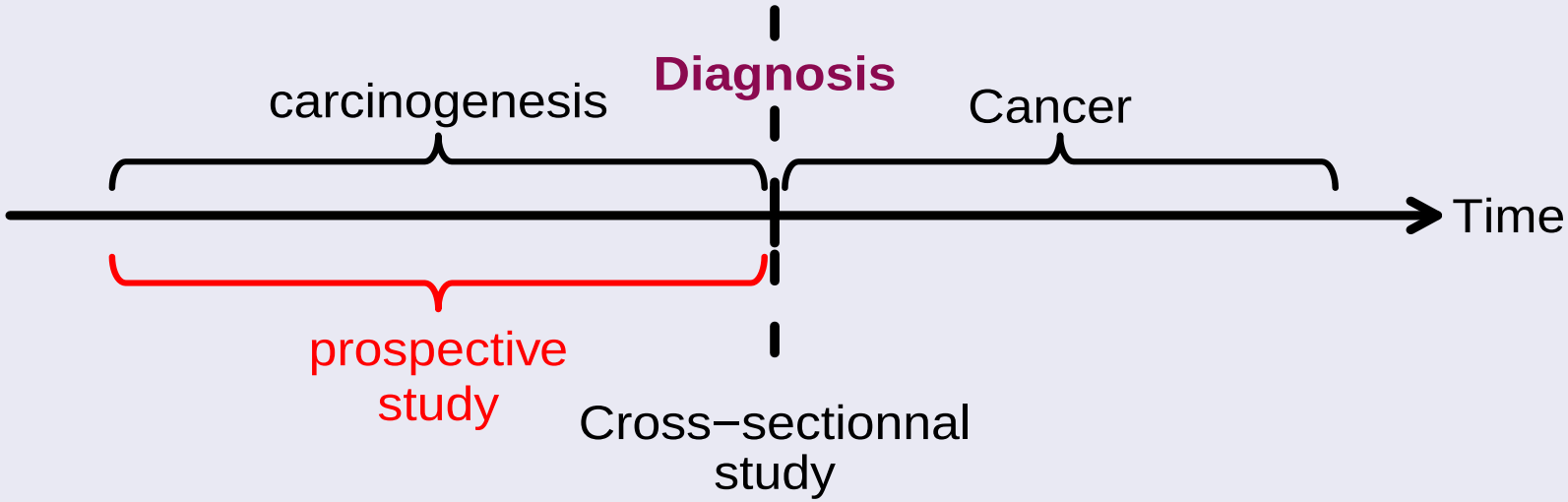
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study

Epidemiology and Carcinogenesis



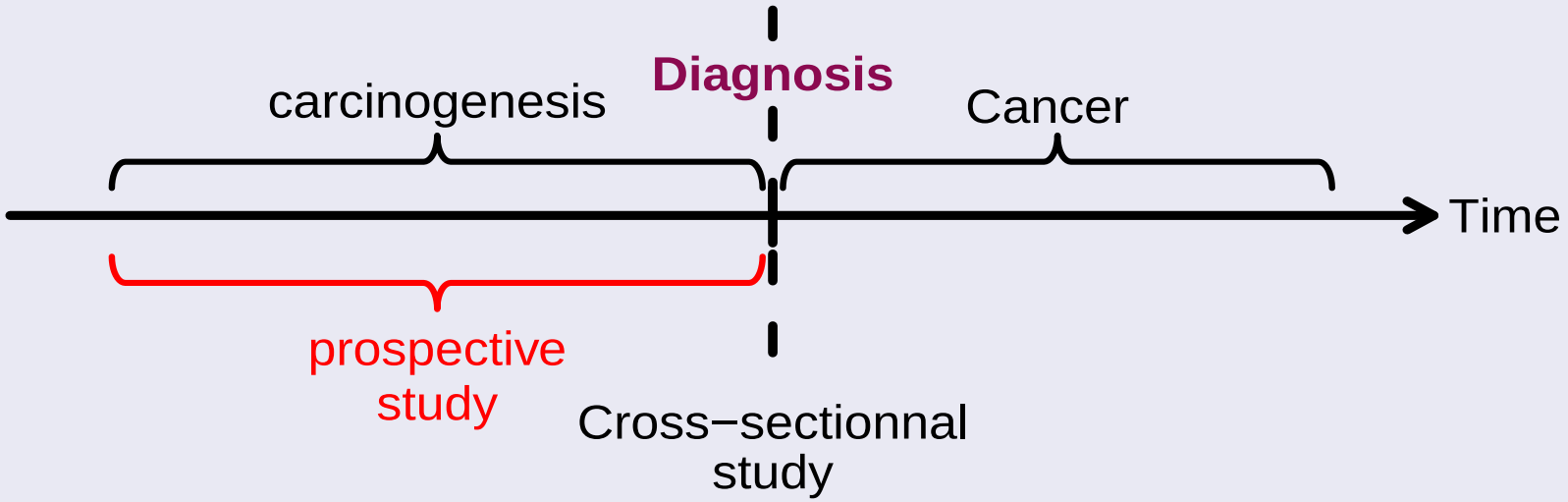
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.

Epidemiology and Carcinogenesis



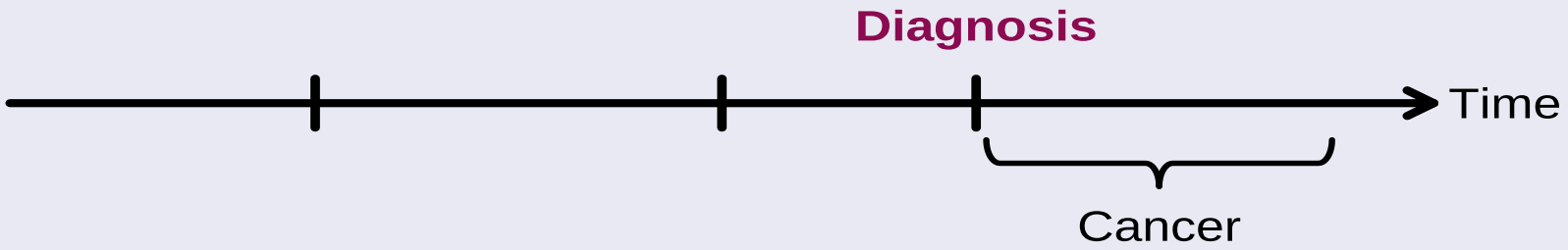
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  ↪ New statistical approach.

Epidemiology and Carcinogenesis

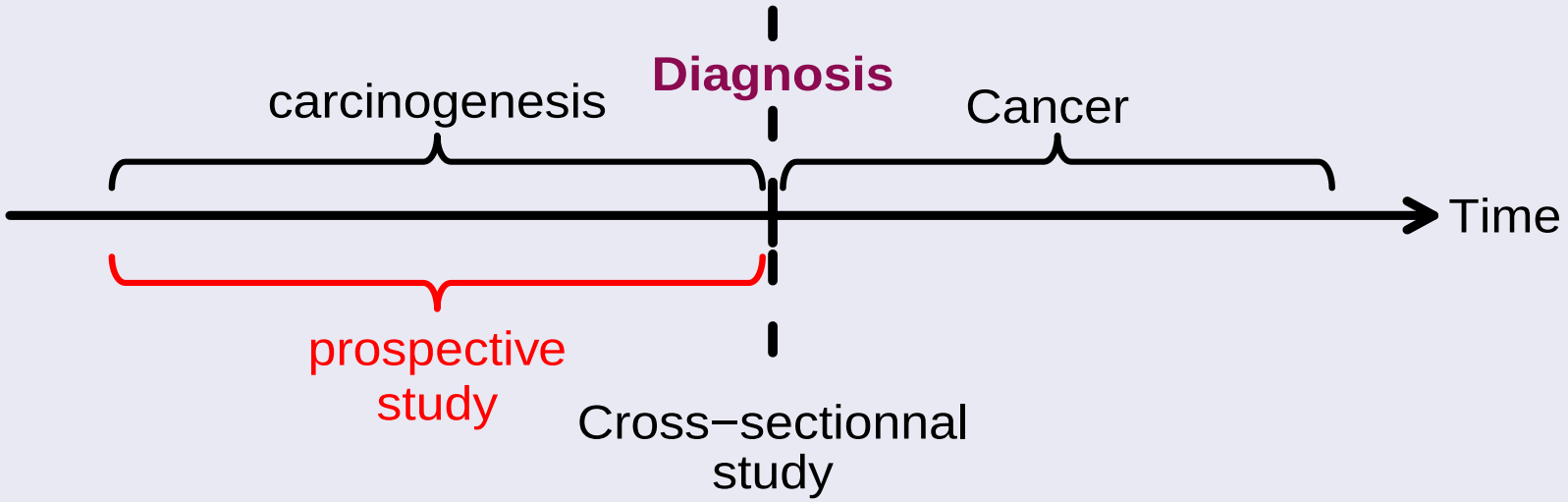


- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  $\hookrightarrow$ New statistical approach.

The multi-stage model

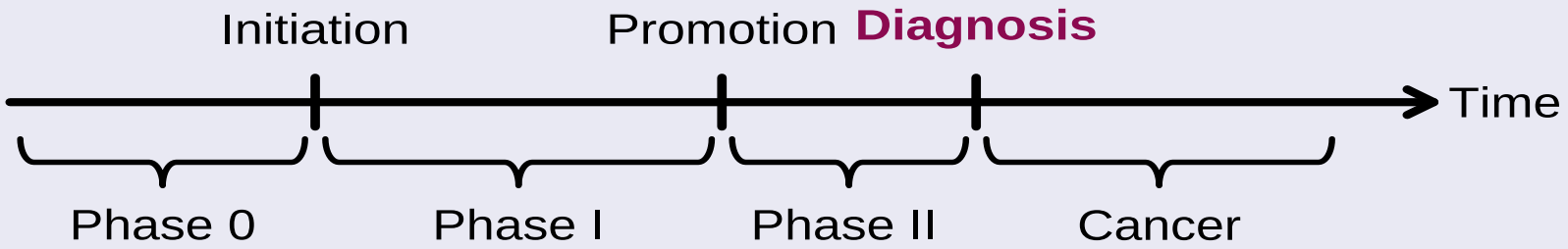


Epidemiology and Carcinogenesis

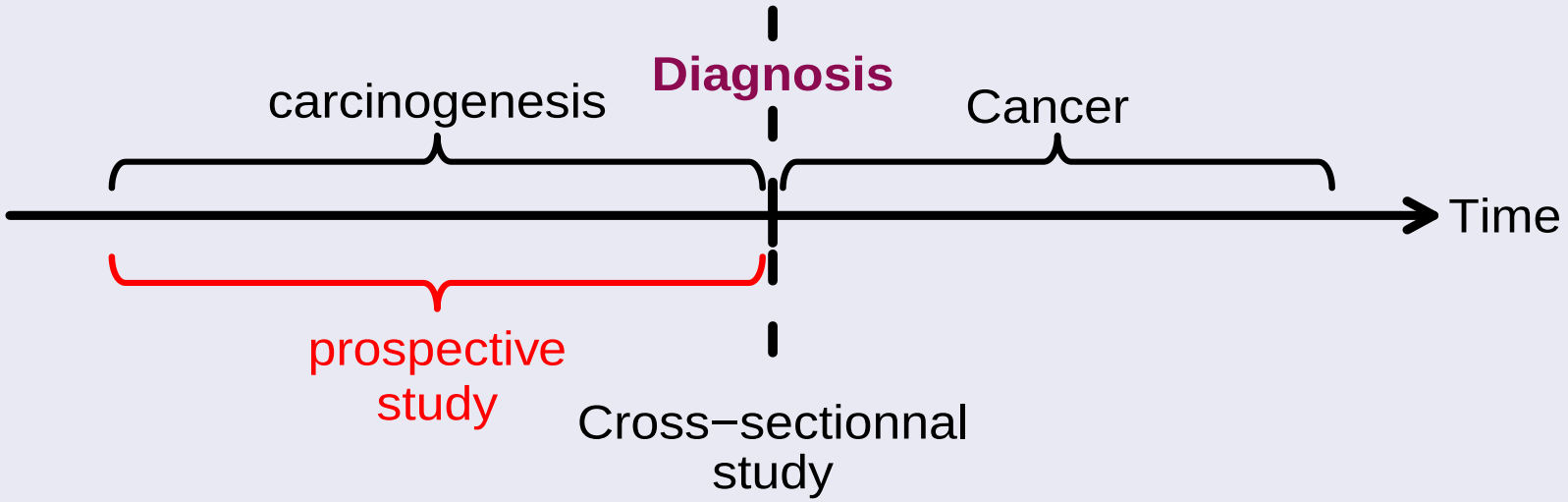


- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  $\hookrightarrow$ New statistical approach.

The multi-stage model

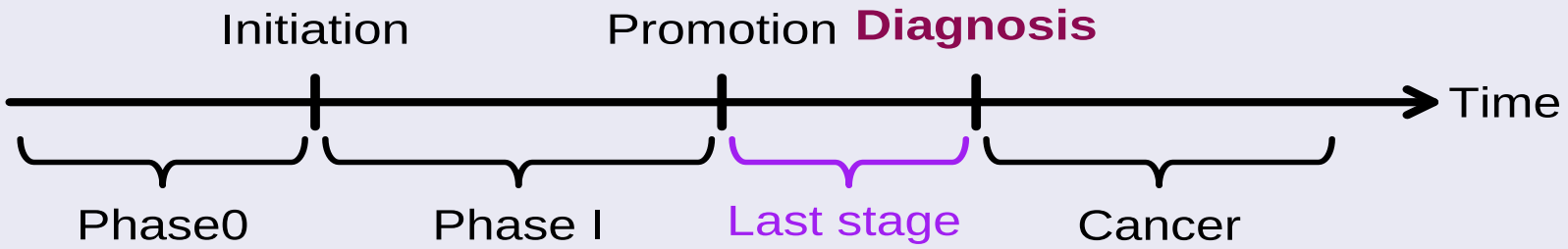


Epidemiology and Carcinogenesis



- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  ↪ New statistical approach.

The multi-stage model



Different levels of statistical modeling in Epidemiology

Different levels of statistical modeling in Epidemiology

- **General statistical methods:** survival analysis (e.g. Cox), gene-by-gene tests.

  $\hookrightarrow$ No biological assumption.

Different levels of statistical modeling in Epidemiology

- **General statistical methods:** survival analysis (e.g. Cox), gene-by-gene tests.

  $\hookrightarrow$ No biological assumption.

- **Causal modeling:** complex system approach.

  $\hookrightarrow$ Recently developed

  $\hookrightarrow$ Precise parametrization of biological/epidemiological phenomenons.

  $\hookrightarrow$ Use of prior information

Different levels of statistical modeling in Epidemiology

- **General statistical methods:** survival analysis (e.g. Cox), gene-by-gene tests.

  ↪ No biological assumption.

- **Our approach:**

  ↪ no causal modeling.

  ↪ model of gene expression evolution during carcinogenesis.

- **Causal modeling:** complex system approach.

  ↪ Recently developed

  ↪ Precise parametrization of biological/epidemiological phenomenons.

  ↪ Use of prior information

Different levels of statistical modeling in Epidemiology

- **General statistical methods:** survival analysis (e.g. Cox), gene-by-gene tests.

  ↪ No biological assumption.

- **Our approach:**

  ↪ no causal modeling.

  ↪ model of gene expression evolution during carcinogenesis.

- **Causal modeling:** complex system approach.

  ↪ Recently developed

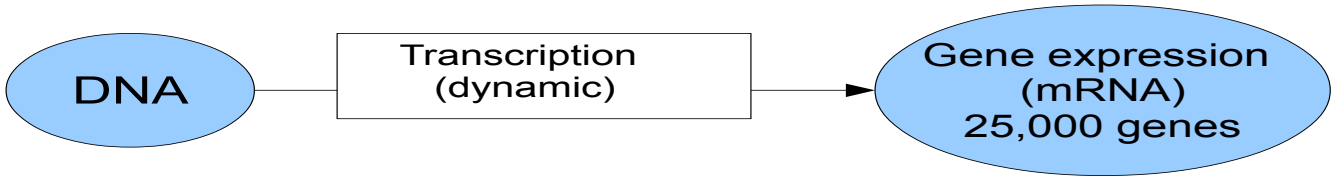
  ↪ Precise parametrization of biological/epidemiological phenomenons.

  ↪ Use of prior information

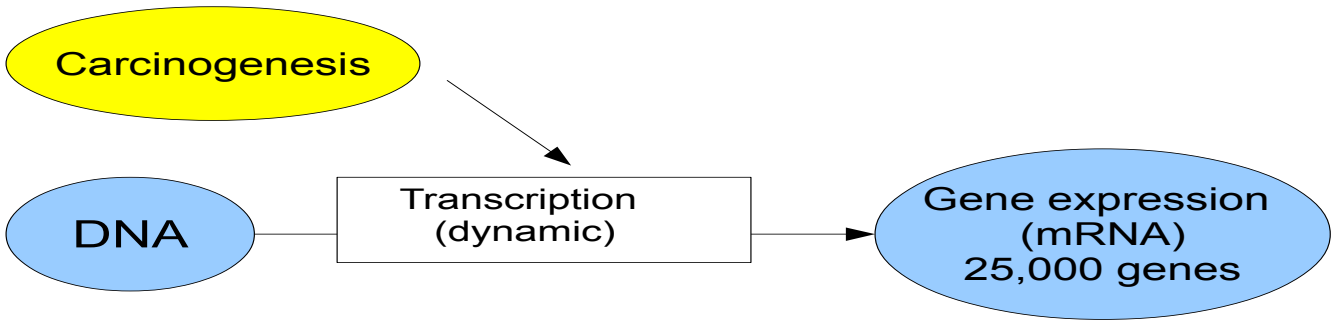
The results from these different approaches can be compared and reinforce/validate the biological model.

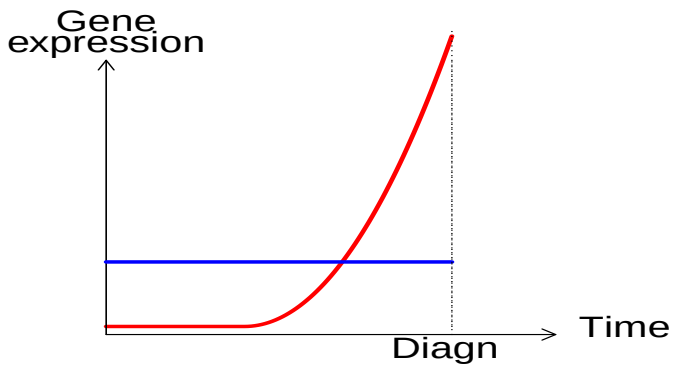
1. Multi-stage model and gene expression
2. The NOWAC data
3. The latent variable statistical model

Transcription



Transcription

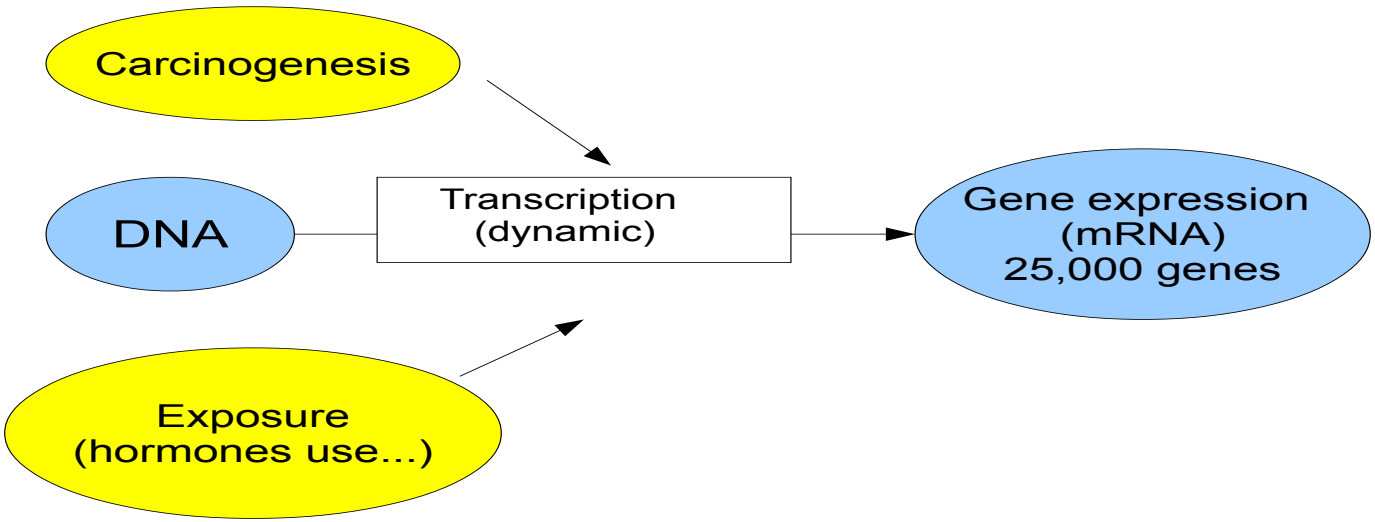


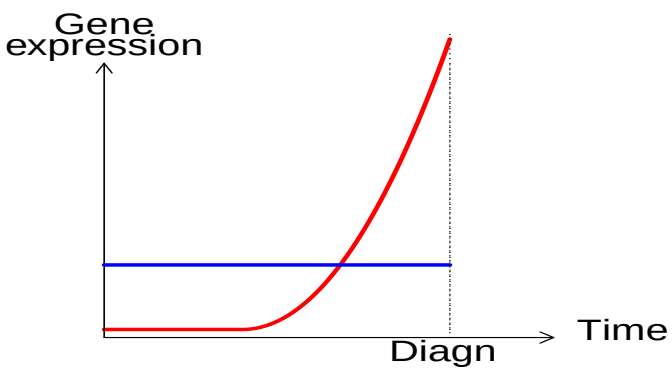


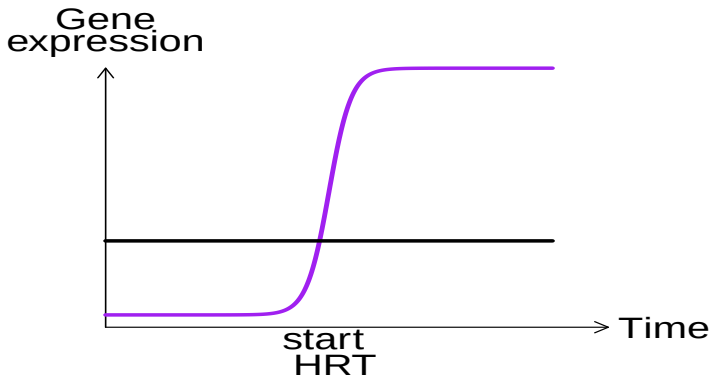
— gene involved in carcinogenesis

— gene non involved

Transcription



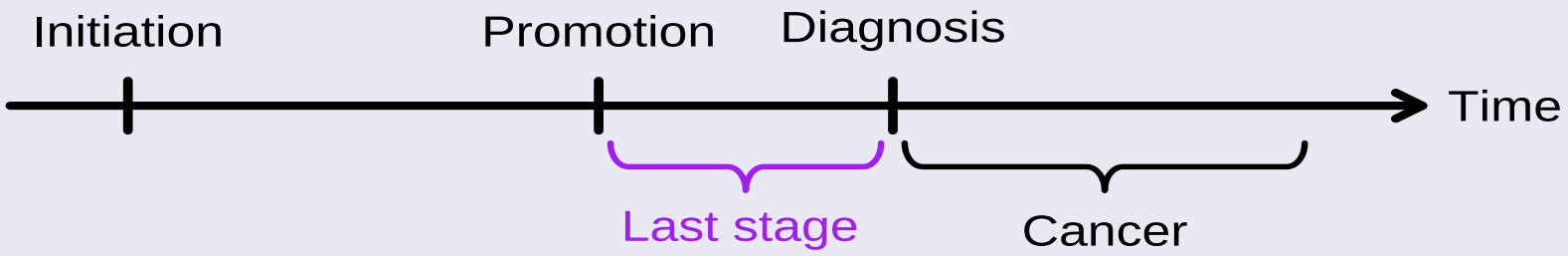




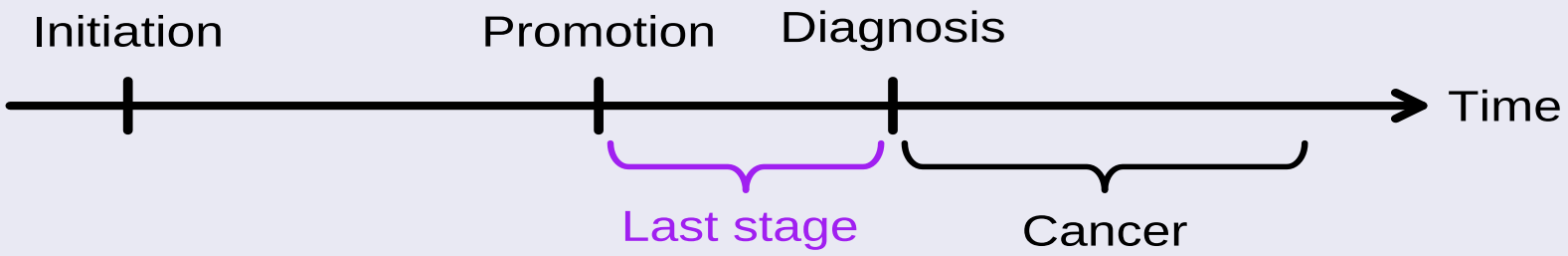
— gene involved in carcinogenesis
— gene non involved

— gene linked to HRT
— gene non-linked to HRT

Multi-stage model and gene expression

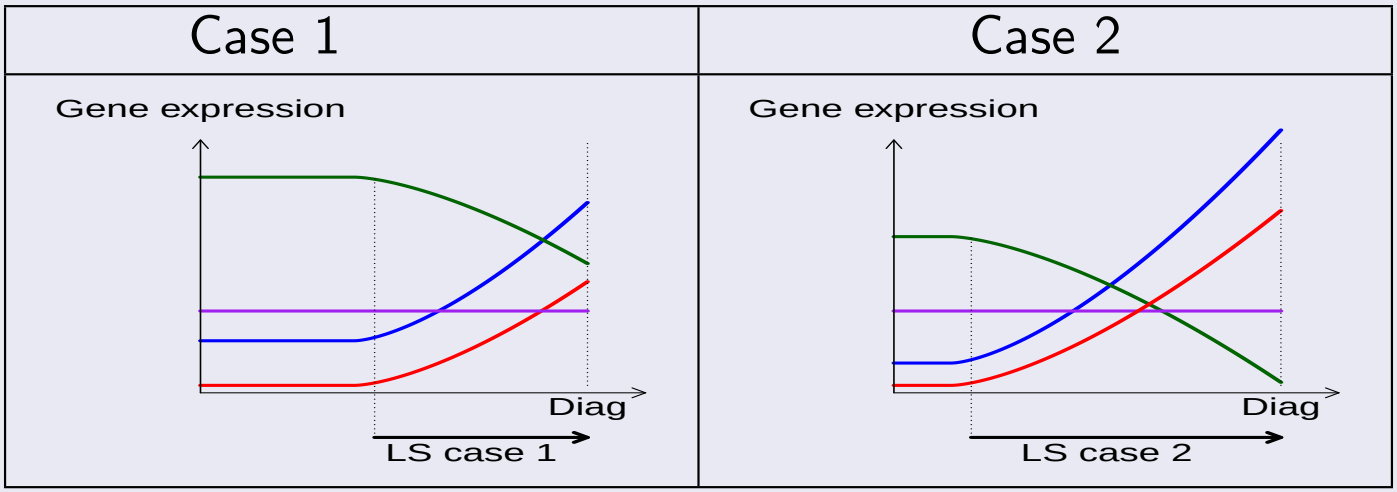


Multi-stage model and gene expression



Multi-stage model and gene expression

- At beginning of last stage, the genes involved in carcinogenesis start to over/under express.
- Random last stage length (=LS)



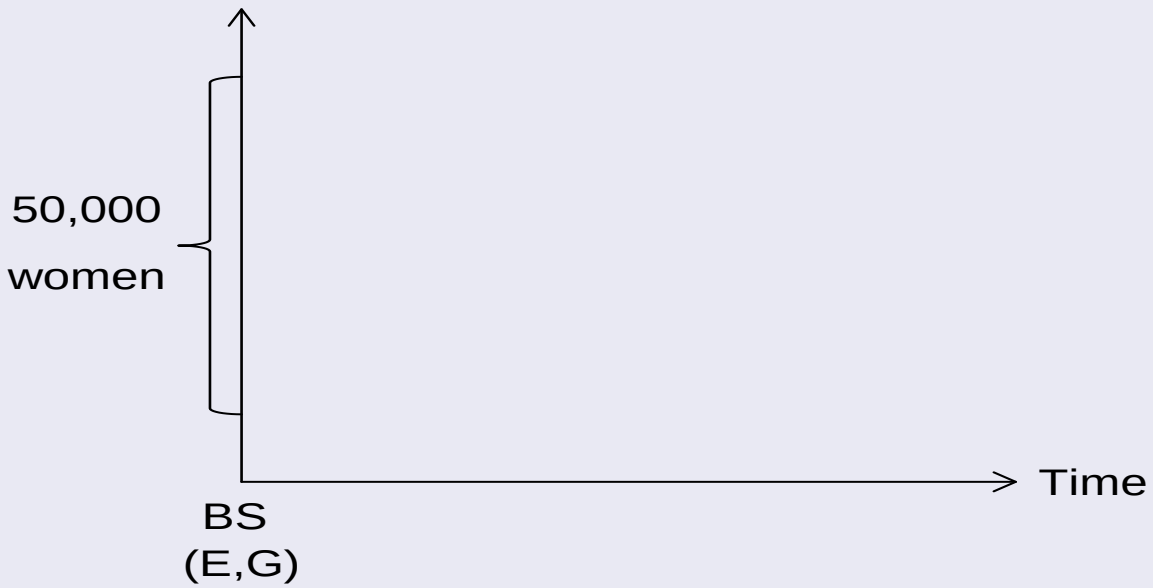
1 Multi-stage model and gene expression

2 The NOWAC data

3 The latent variable statistical model

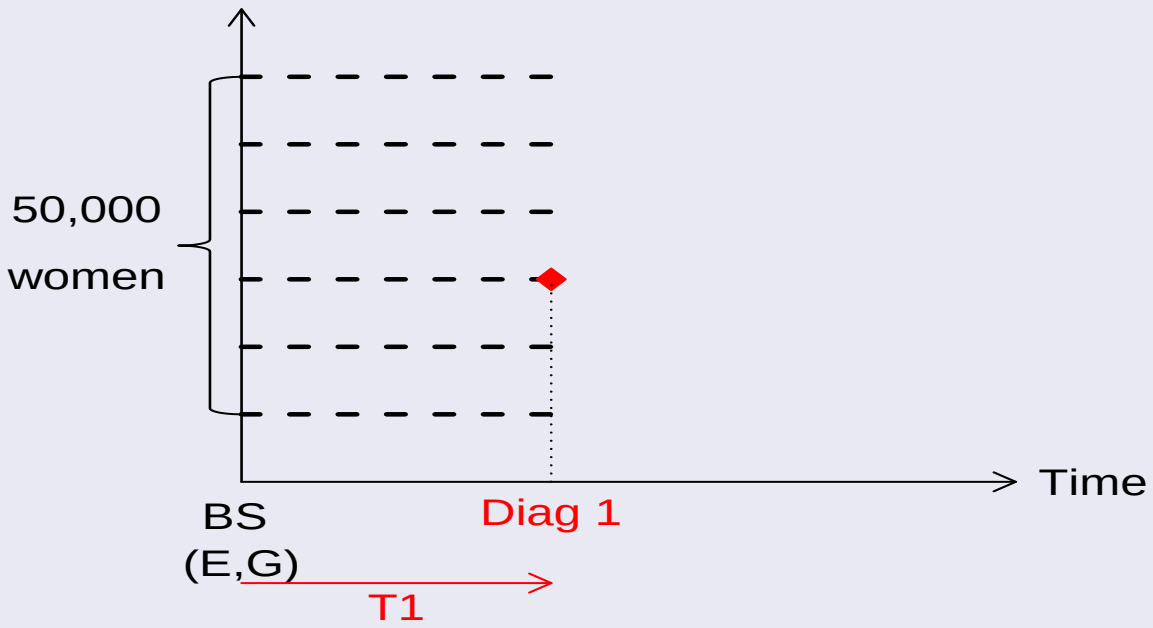
The NOWAC cohort

Cohort of 50,000 women



The NOWAC cohort

Cohort of 50,000 women

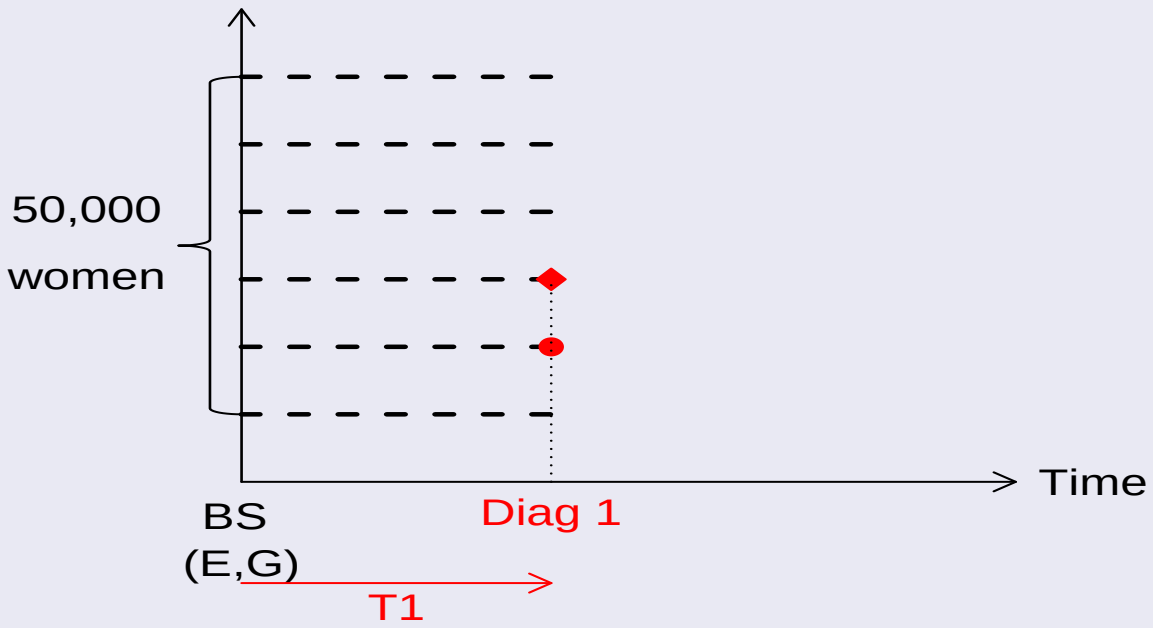


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

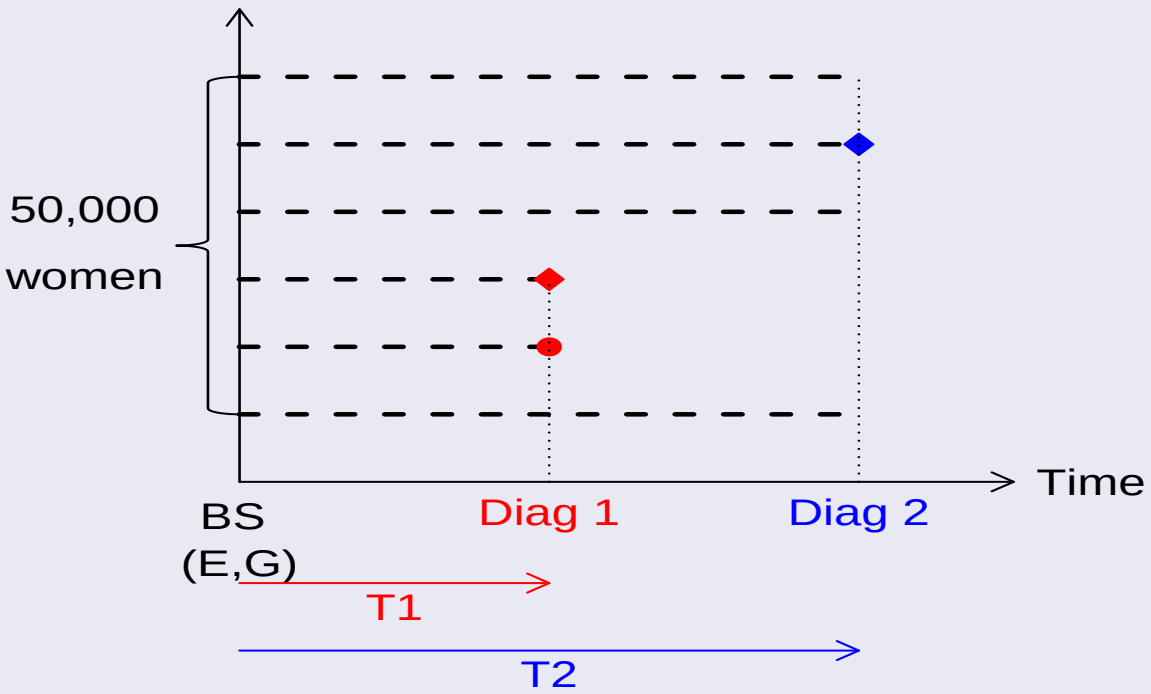


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

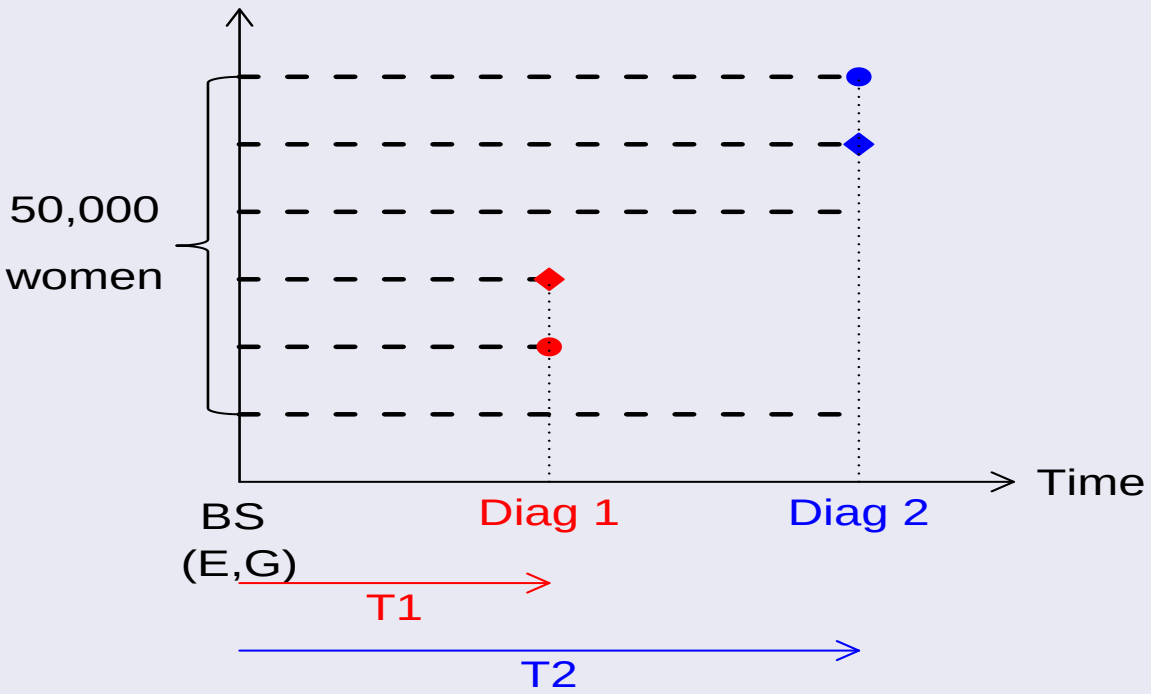


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

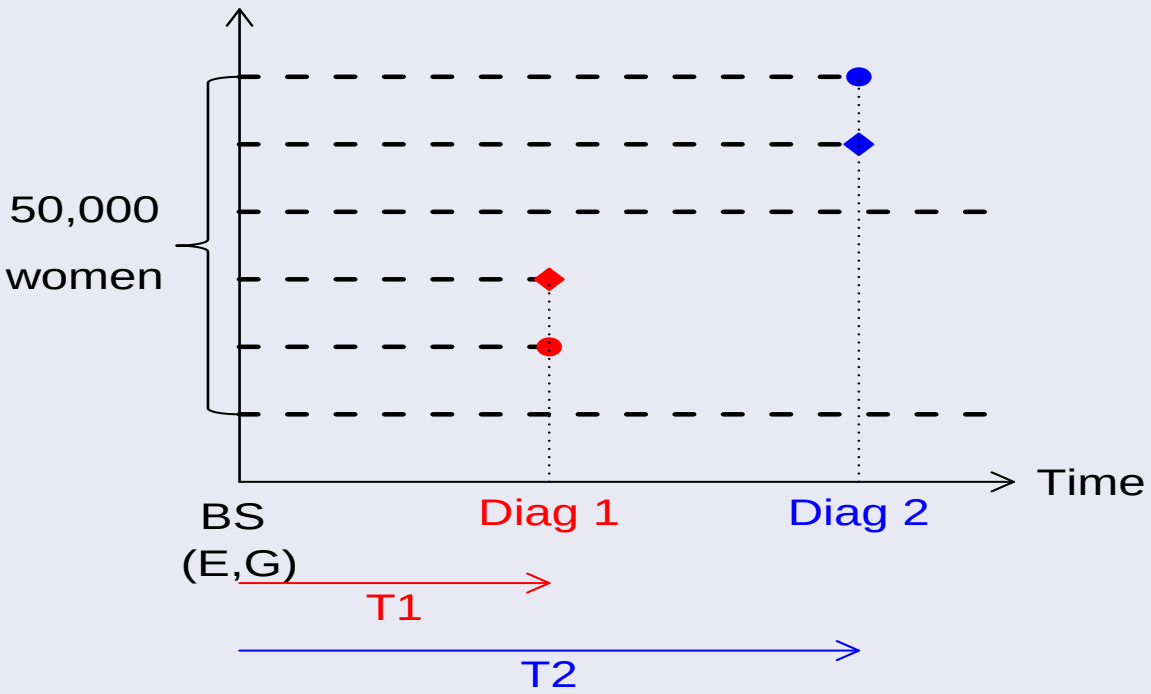


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

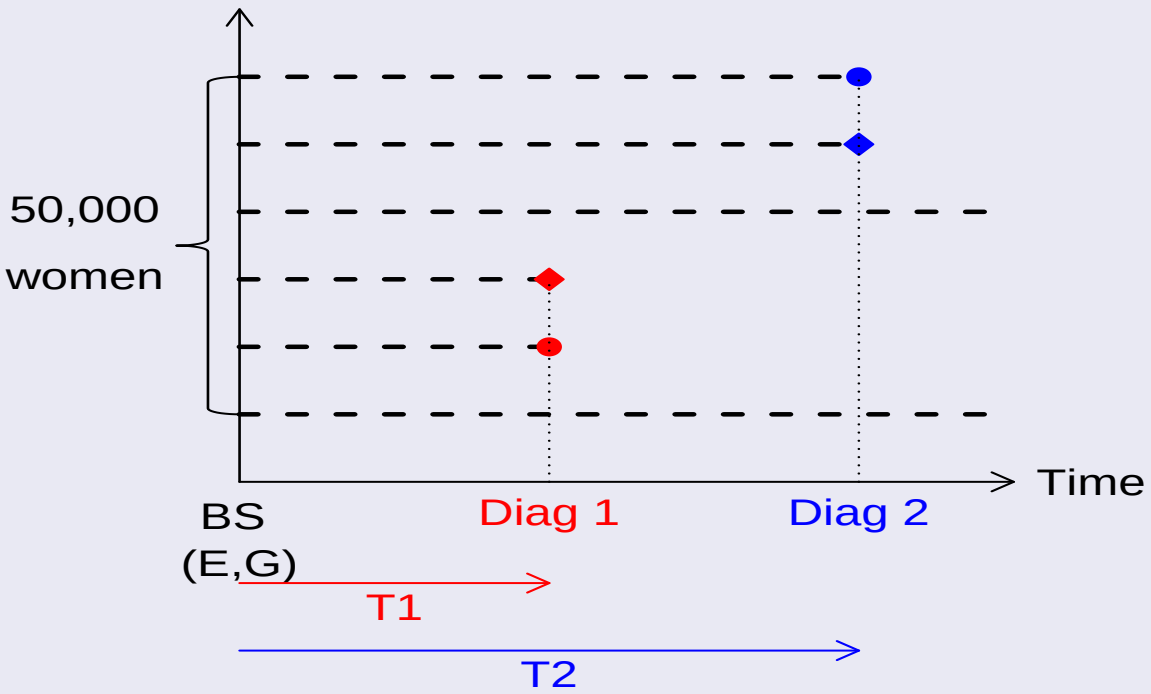


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

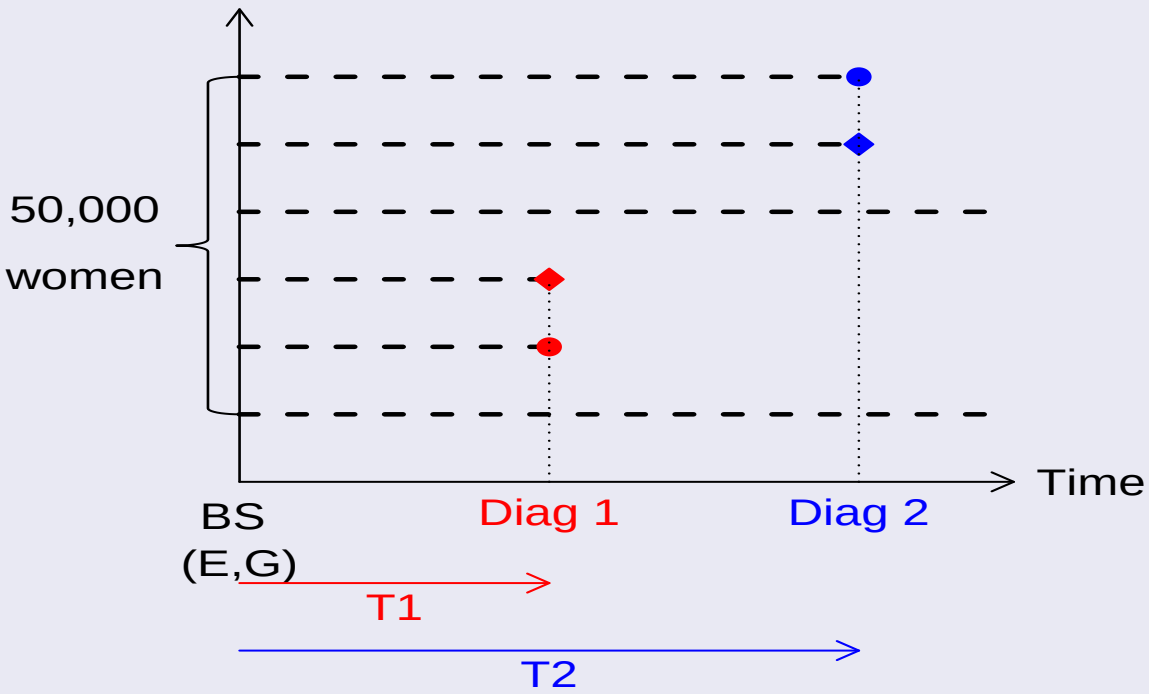


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women



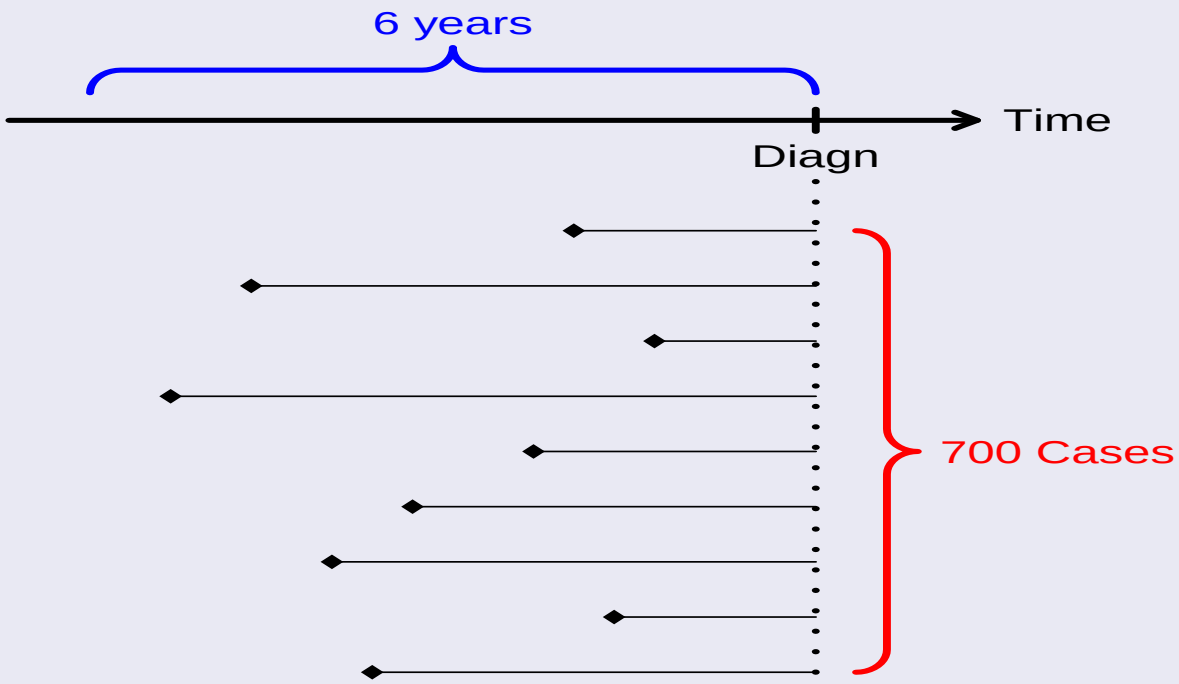
◆: case
●: control

For each case-control pair:

- $(E^{\text{case}}, E^{\text{ctl}})$ = Exposure at time of BS.
- T = Follow-up time.
- DG = Difference of gene expression at time T before diagnosis (25,000 genes).

Set of data

- 6 years of follow-up.
- 700 case-control pairs.
- only one measurement by pair.

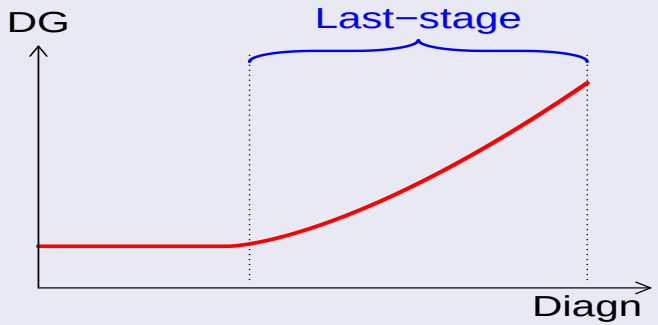


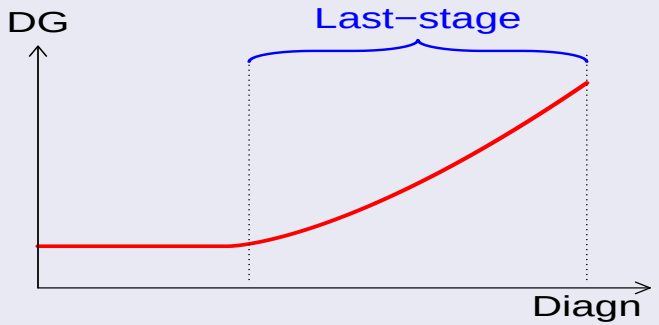
1. Multi-stage model and gene expression
2. The NOWAC data
3. The latent variable statistical model

.

The latent last-stage statistical model

Gene expression

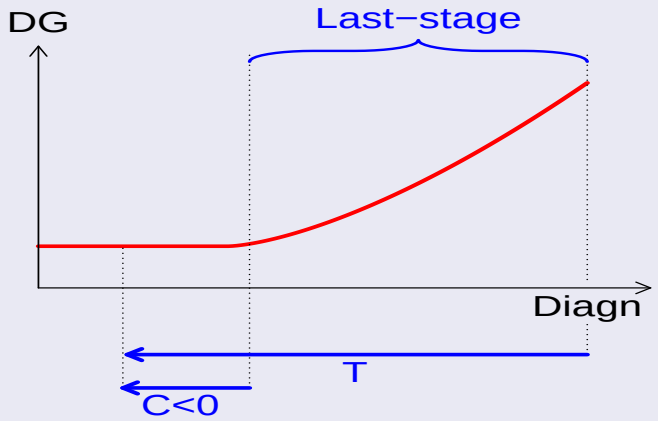


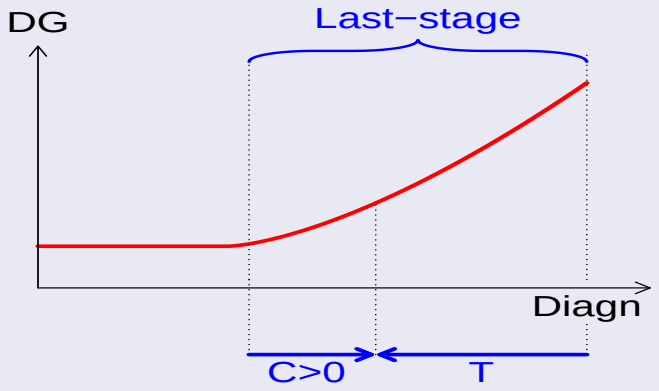


- Linear dependence on time.

The latent last-stage statistical model

Gene expression

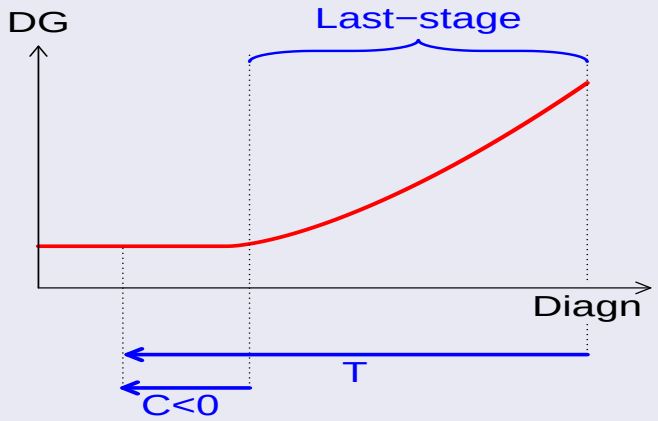


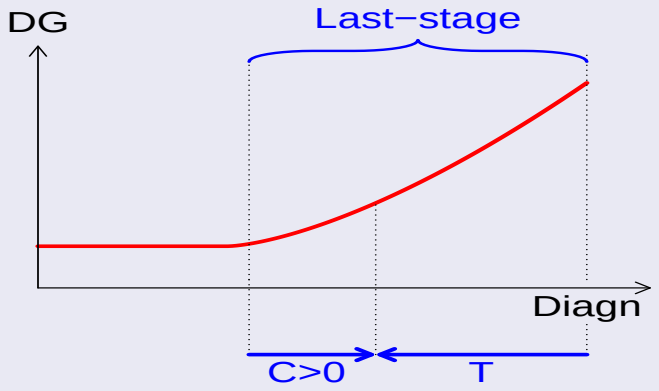


- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$

The latent last-stage statistical model

Gene expression

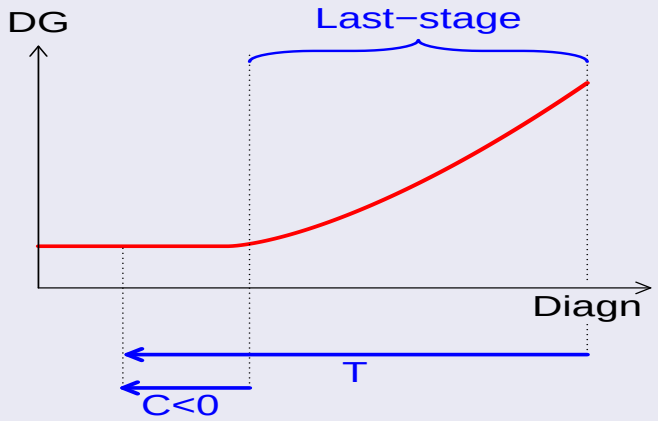


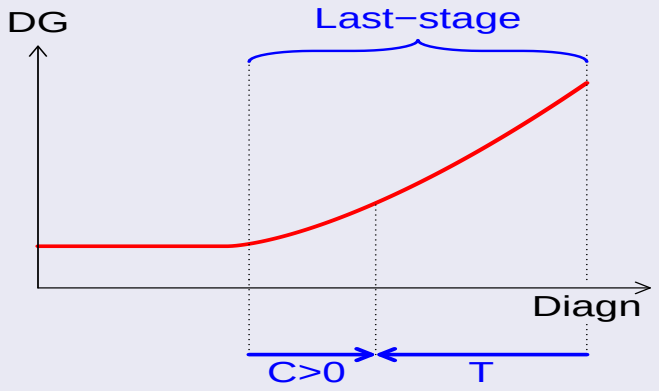


- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.

The latent last-stage statistical model

Gene expression

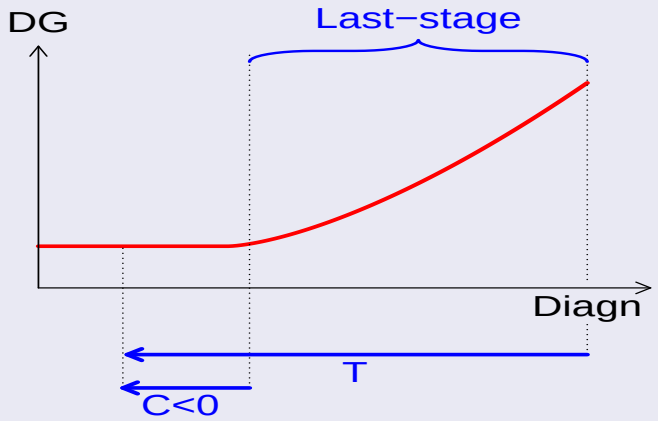


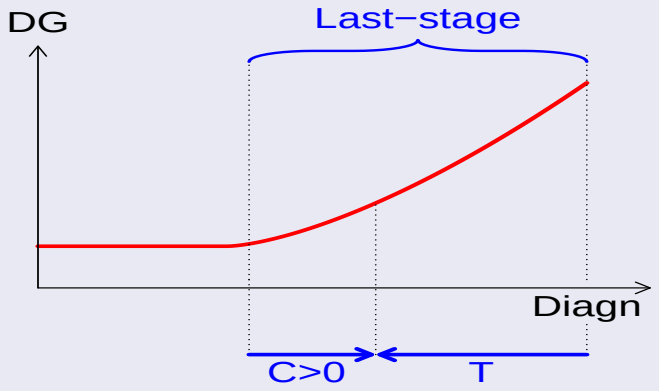


- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.
- Let DE be the "difference of exposures" between case and control.

The latent last-stage statistical model

Gene expression

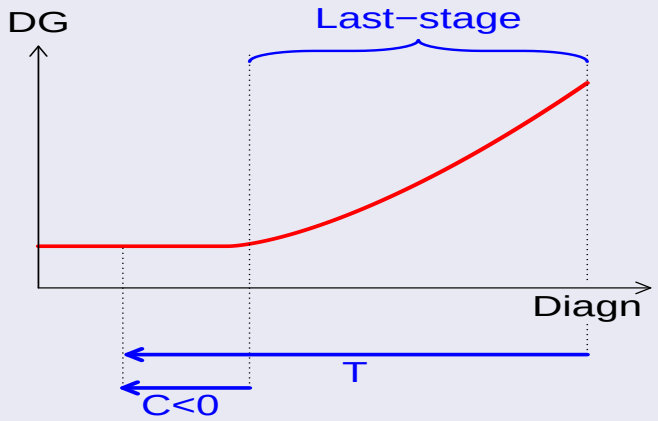


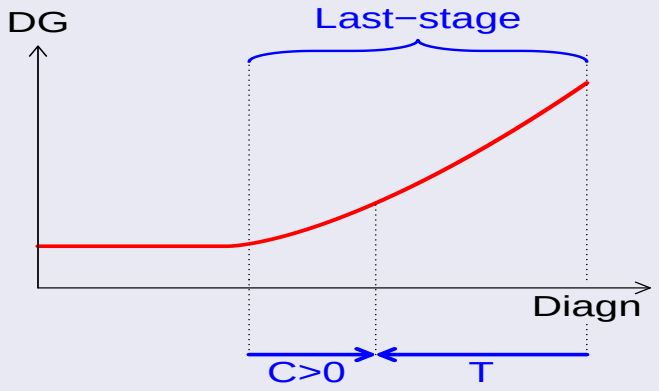


- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.
- Let DE be the "difference of exposures" between case and control.
- For each gene g, $DG^g = \beta_0^g + \langle \beta_1^g, DE \rangle + \beta_2^g C \mathbb{I}(C > 0) + \varepsilon_g$

The latent last-stage statistical model

Gene expression



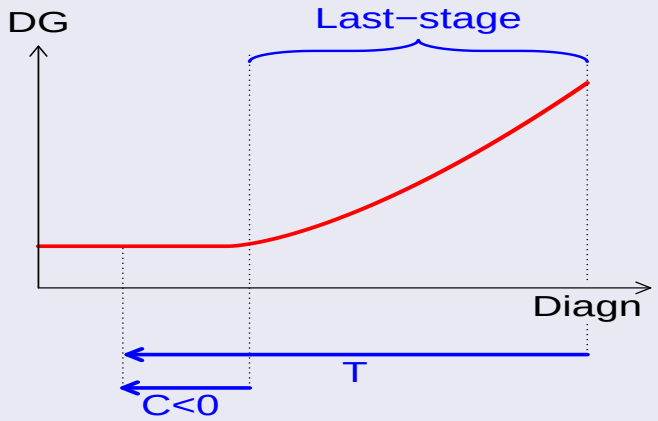


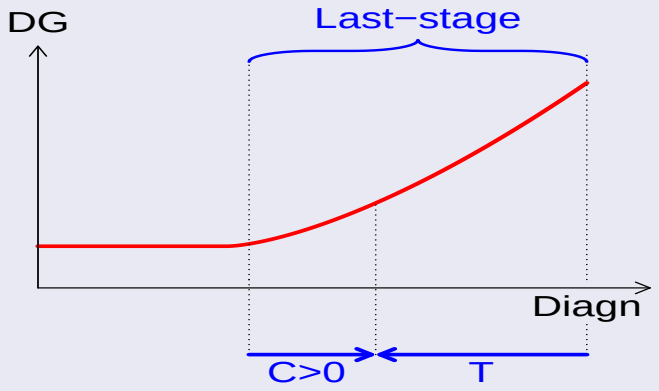
- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.
- Let DE be the "difference of exposures" between case and control.
- For each gene g, $DG^g = \beta_0^g + \langle \beta_1^g, DE \rangle + \beta_2^g C \mathbb{I}(C > 0) + \varepsilon_g$

and g is involved in the last-stage iif $\beta_2^g \neq 0$.

The latent last-stage statistical model

Gene expression





- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.
- Let DE be the "difference of exposures" between case and control.
- For each gene g, $DG^g = \beta_0^g + \langle \beta_1^g, DE \rangle + \beta_2^g C \mathbb{I}(C > 0) + \varepsilon_g$

and g is involved in the last-stage iif $\beta_2^g \neq 0$.

Last-stage length

$LS \sim \Gamma(k, \theta)$, with (k, θ) dependent on the exposures of the case E^{case}

Estimation of the model

- Algorithm SEM (Simulated Expectation Maximization)
- Validation on simulated data

Estimation of the model

- Algorithm SEM (Simulated Expectation Maximization)
- Validation on simulated data

Primary goals

- Detect genes involved in the last stage (multiple testing).
- Estimate the distribution of the last stage depending on the exposures.

Further developments

Further developments

Modeling of $P[DG|C, DE]$

Further developments

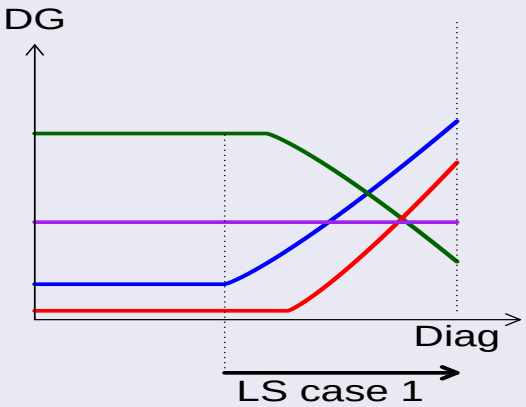
Modeling of $P[DG|C, DE]$

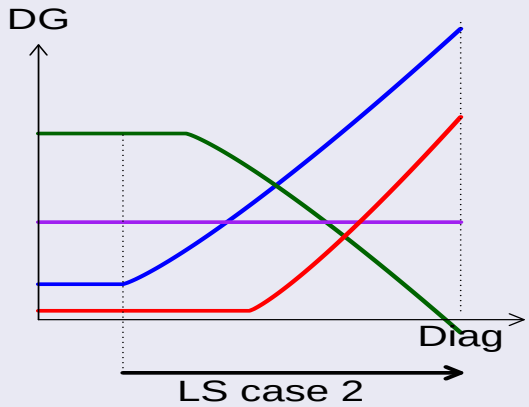
- Alternative parametrization of time dependence.

Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.

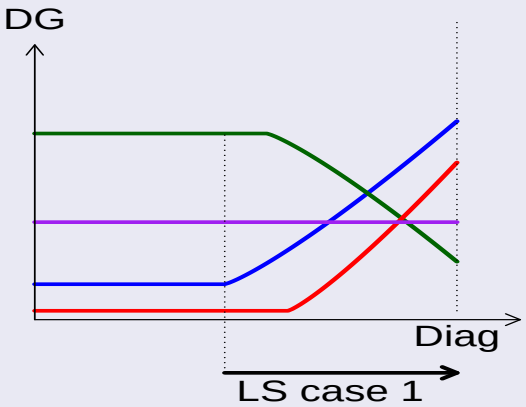


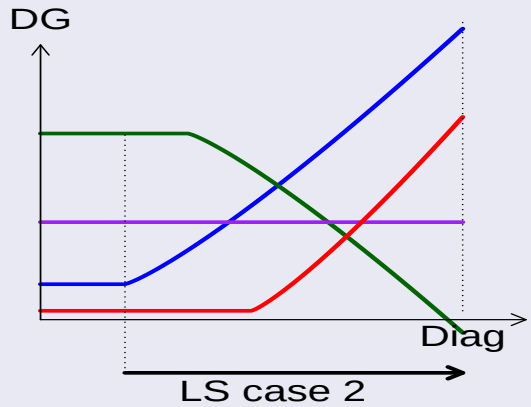


Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.



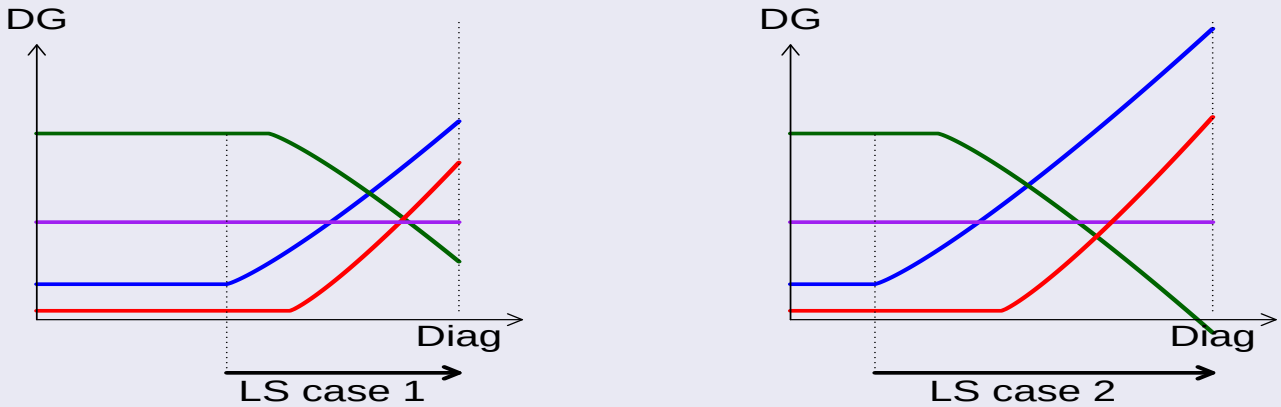


- Cross-effect Time/Exposure (Cancer driven by exposure).

Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.

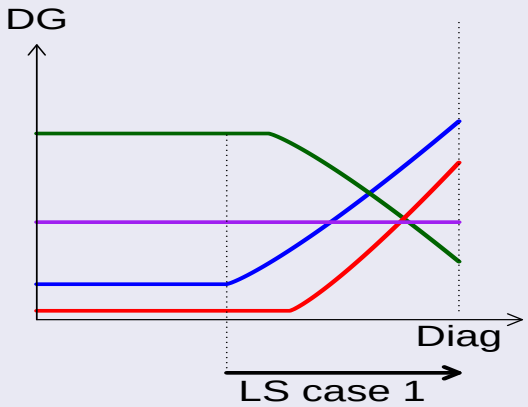


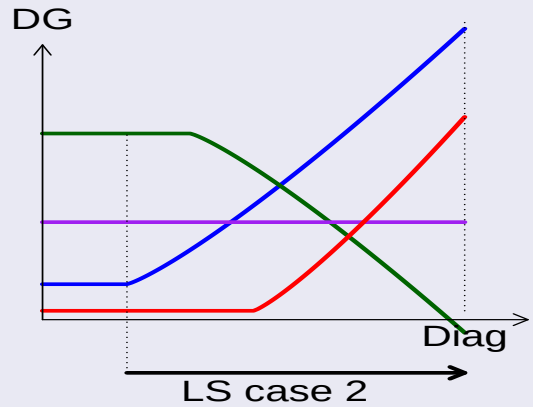
- Cross-effect Time/Exposure (Cancer driven by exposure).
- Dependence between genes: - Statistical inference
  - A priori knowledge (Gene Ontology,...)

Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.





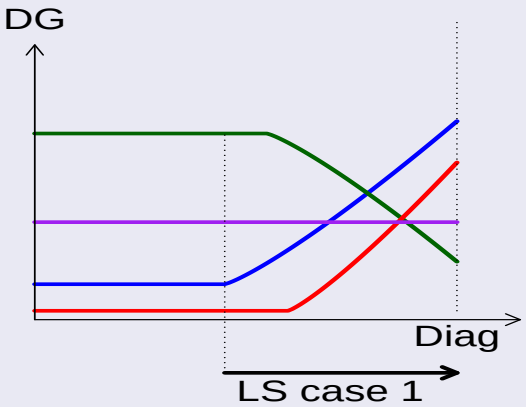
- Cross-effect Time/Exposure (Cancer driven by exposure).
- Dependence between genes: - Statistical inference
  - A priori knowledge (Gene Ontology,...)

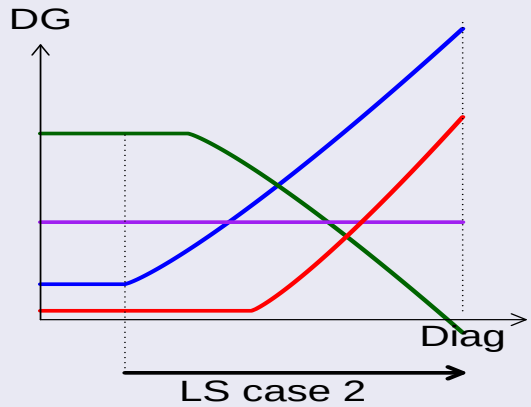
Subgroups

Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.





- Cross-effect Time/Exposure (Cancer driven by exposure).
- Dependence between genes: - Statistical inference
  - A priori knowledge (Gene Ontology,...)

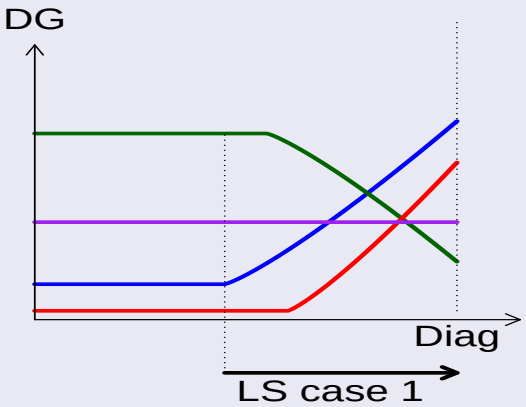
Subgroups

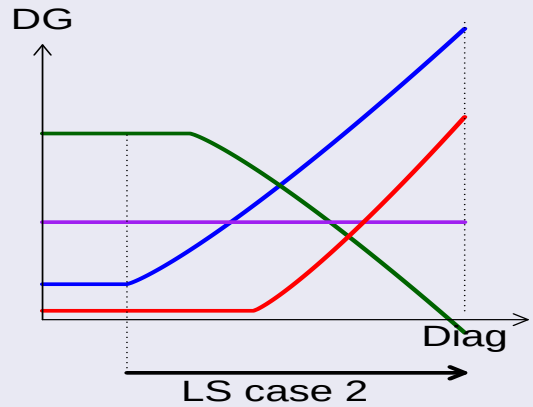
- Classification by exposure/type of cancer

Further developments

Modeling of $P[DG|C, DE]$

- Alternative parametrization of time dependence.
- Interval between between beginning of last stage and gene expression changes.





- Cross-effect Time/Exposure (Cancer driven by exposure).
- Dependence between genes: - Statistical inference
  - A priori knowledge (Gene Ontology,...)

Subgroups

- Classification by exposure/type of cancer
- Stratification/hierarchical model.

Conclusion

- Goal: study gene expression through the multi-stage model of carcinogenesis.
- Conceptual model based on biological modeling
- Good results on simulations but require development to be applied on experimental data.