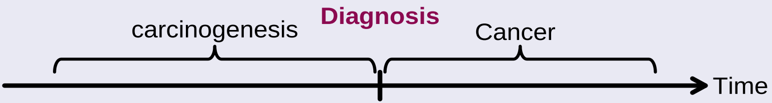
A statistical approach for carcinogenesis in transcriptomics

TICE (Transcriptomics In Cancer Epidemiology)
NOWAC (Norwegian Women And Cancer)

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

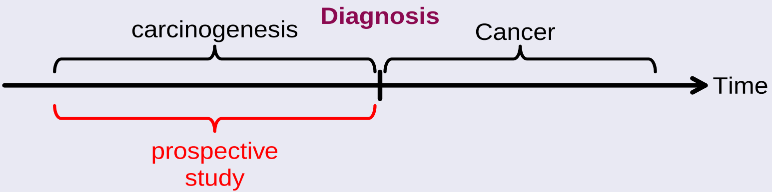
22th of May 2012

Epidemiology and Carcinogenesis



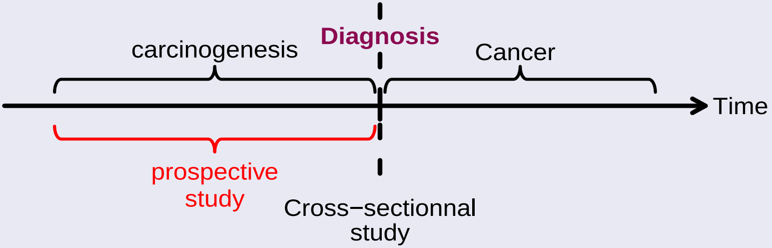
- Carcinogenesis: prior to diagnosis

Epidemiology and Carcinogenesis



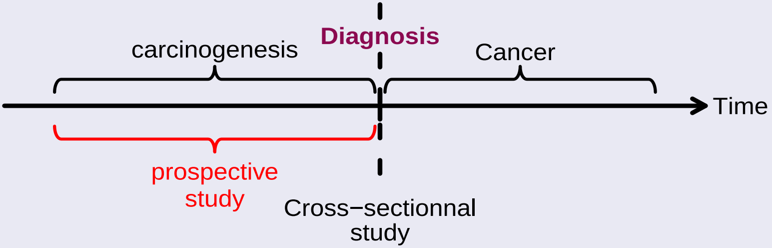
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study

Epidemiology and Carcinogenesis



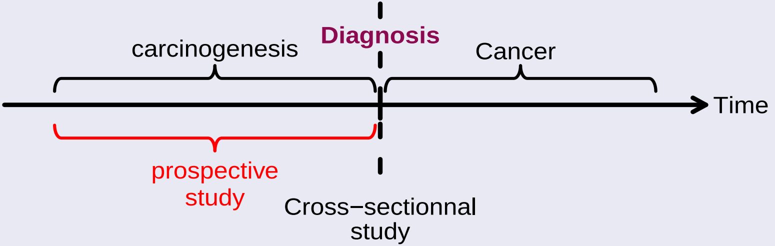
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.

Epidemiology and Carcinogenesis



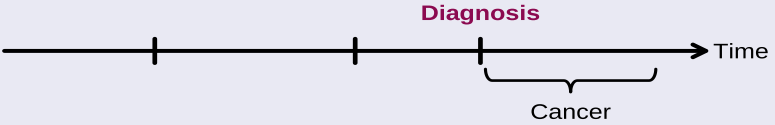
- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  ↪ New statistical approach.

Epidemiology and Carcinogenesis

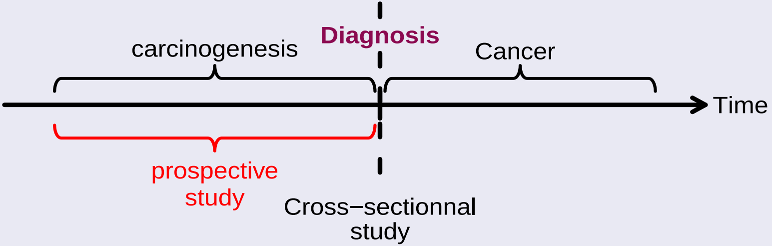


- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  ↪ New statistical approach.

The multi-stage model

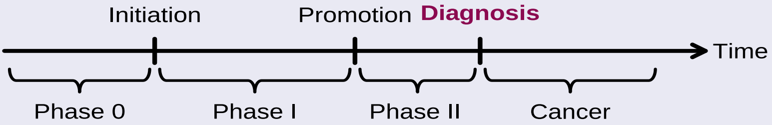


Epidemiology and Carcinogenesis

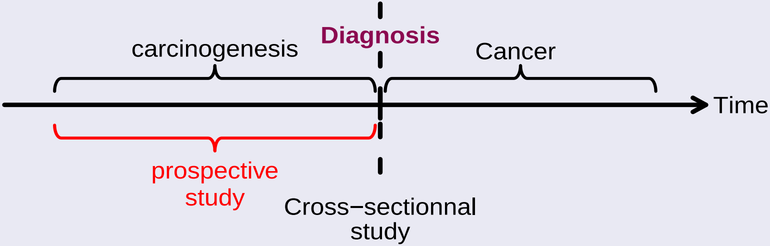


- Carcinogenesis: prior to diagnosis
- NOWAC: Prospective study
- Usually: gene expression studied in cross-sectional design.
  ↪ New statistical approach.

The multi-stage model

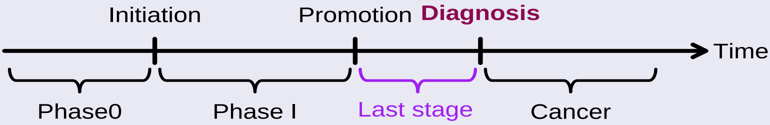


Epidemiology and Carcinogenesis



- 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.

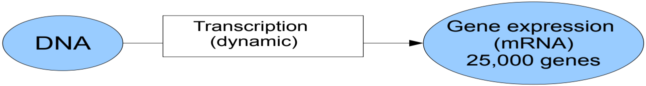
  ↪ Recently developed

  ↪ Precise parametrization of biological/epidemiological phenomenons.

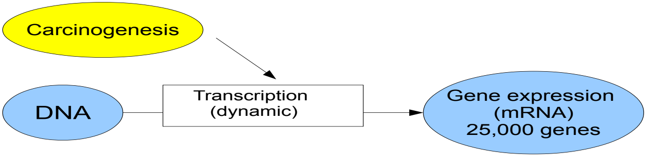
  ↪ Use of prior information

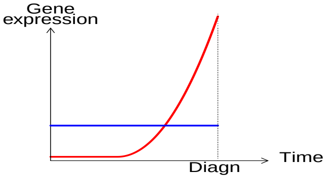
The results from these different approaches can be compared and reinforce/validate the biological model.

Transcription



Transcription

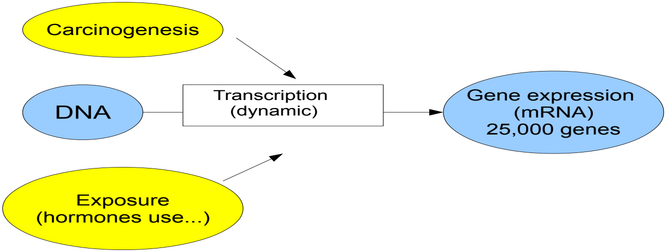


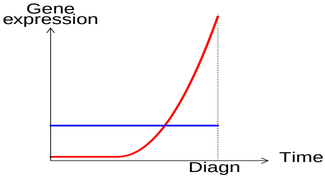


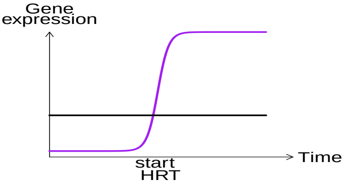
— gene involved in carcinogenesis

— gene non involved

Transcription



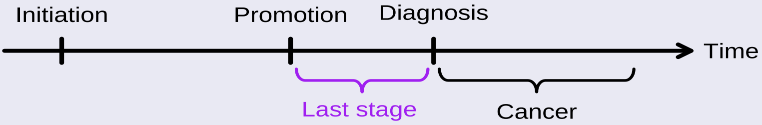




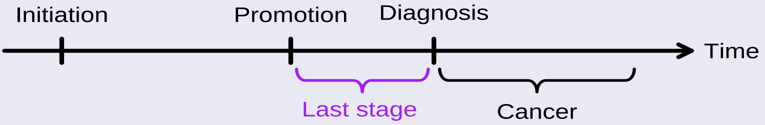
— gene involved in carcinogenesis
— gene non involved

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

Multi-stage model and gene expression

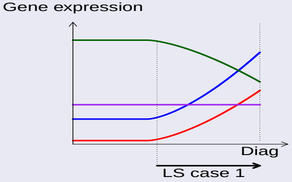
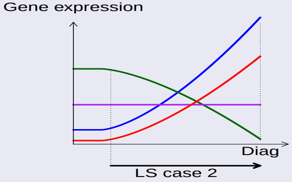


Multi-stage model and gene expression



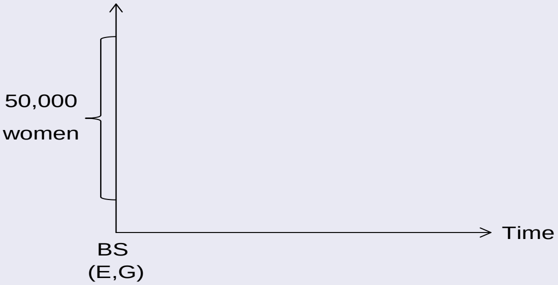
Multi-stage model and gene expression

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

Case 1	Case 2
	

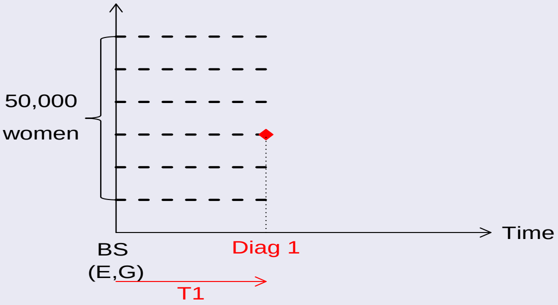
The NOWAC cohort

Cohort of 50,000 women



The NOWAC cohort

Cohort of 50,000 women

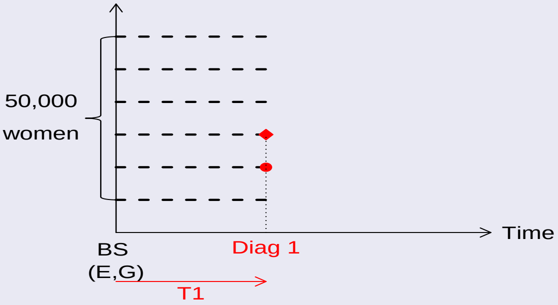


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

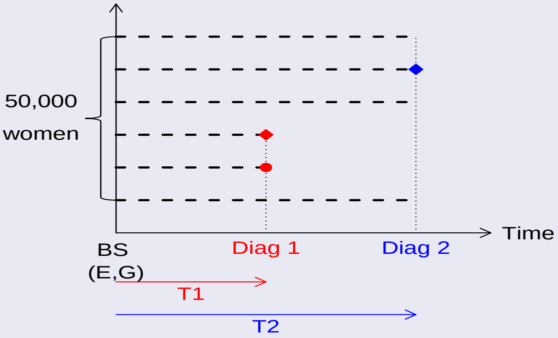


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

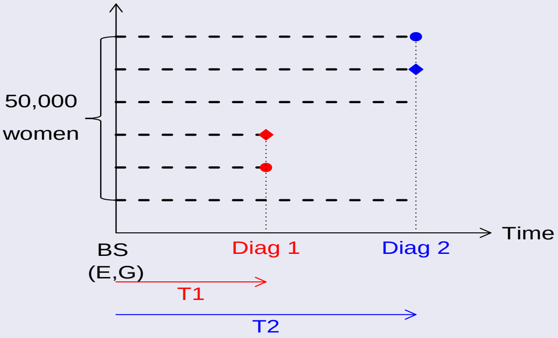


◆: case

●: control

The NOWAC cohort

Cohort of 50,000 women

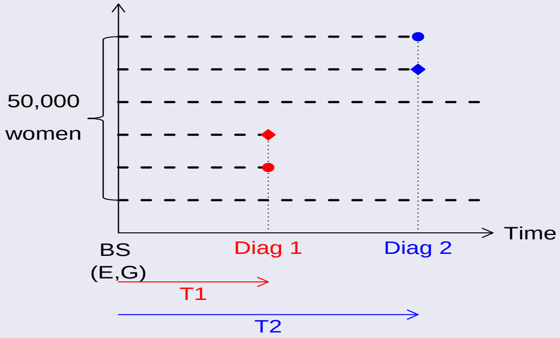


◆: case

●: control

The NOWAC cohort

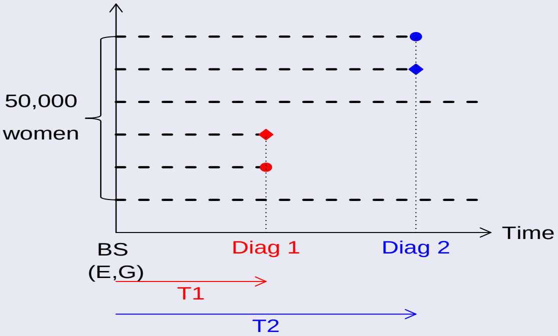
Cohort of 50,000 women



◆: case
●: control

The NOWAC cohort

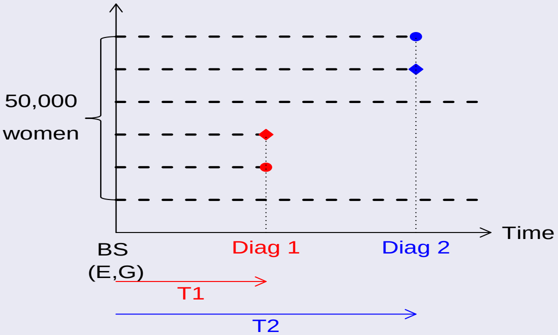
Cohort of 50,000 women



◆: case
●: control

The NOWAC cohort

Cohort of 50,000 women



◆: case

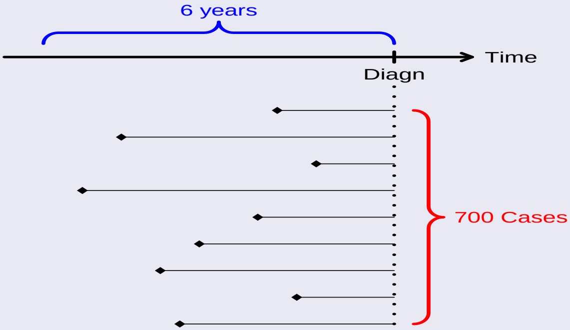
●: control

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).

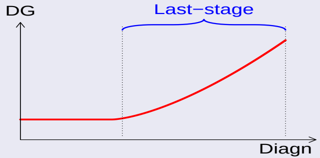
Set of data

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



The latent last-stage model

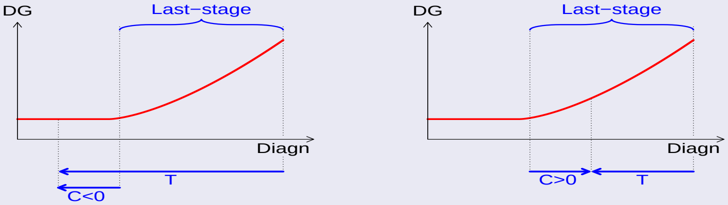
Gene expression



- Linear dependence on time.

The latent last-stage model

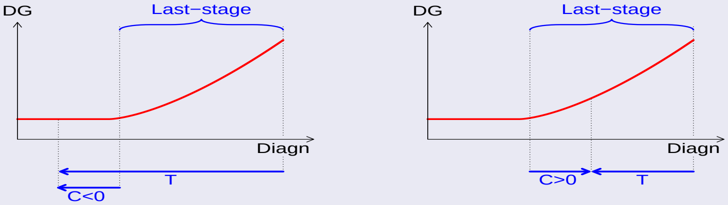
Gene expression



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

The latent last-stage model

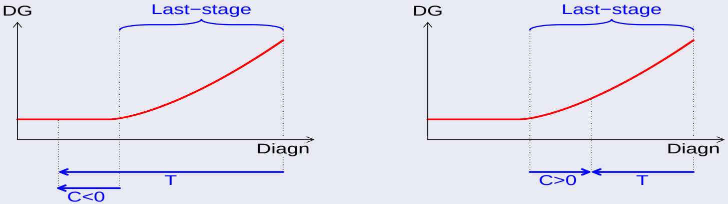
Gene expression



- 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 model

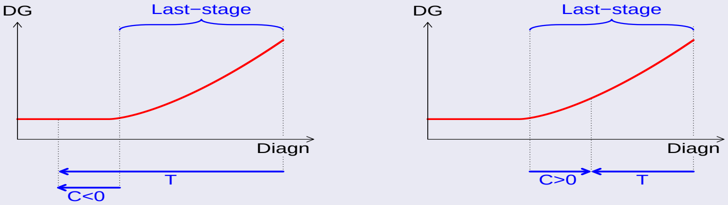
Gene expression



- 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 model

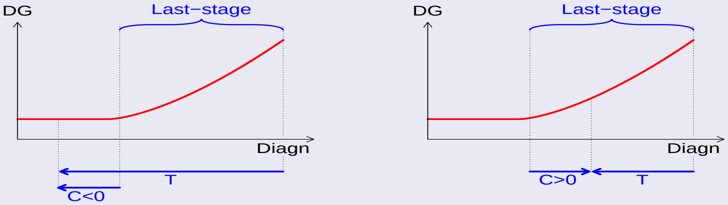
Gene expression



- 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 model

Gene expression

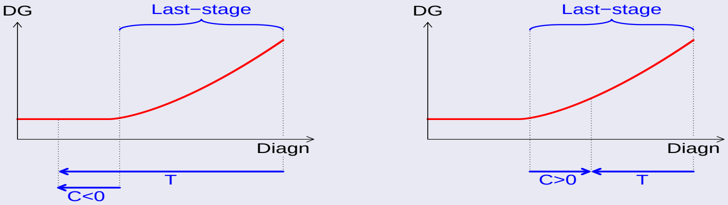


- Linear dependence on time.
- Let $C = LS - T$, DG depends on C iif $C > 0$
- The genes can be constantly differentially expressed before last stage.
- Let DE be the "difference of exposures" between case and control.
- For each gene g, $DG^g = \beta_0^g + \langle \beta_1^g, DE \rangle + \beta_2^g C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

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

The latent last-stage 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}

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

- E^{case}: exposures which affects carcinogenesis
  DE: exposures which affects gene expression.

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

- E^{case}: exposures which affects carcinogenesis
  DE: exposures which affects gene expression.
- We model $P[G|T, E]$.

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

- E^{case}: exposures which affects carcinogenesis
  DE: exposures which affects gene expression.
- We model $P[G|T, E]$.
- Survival analysis model $P[T|G, E]$.

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

- E^{case}: exposures which affects carcinogenesis
  DE: exposures which affects gene expression.
- We model $P[G|T, E]$.
- Survival analysis model $P[T|G, E]$.
  $\rightarrow$ difficult to interpret when G depends on T.

Some comments about the model

Model

For a case-control pair i, $LS_i = C_i + T_i \sim \Gamma(k, \theta)$ where

$$\begin{cases} k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \\ \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \end{cases}$$

For each gene g, $DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}$ where $\varepsilon \sim \mathcal{N}(0, \sigma_g)$.

- E^{case}: exposures which affects carcinogenesis
  DE: exposures which affects gene expression.
- We model $P[G|T, E]$.
- Survival analysis model $P[T|G, E]$.
  $\rightarrow$ difficult to interpret when G depends on T.
- Two main goals:
  - Estimate last-stage length distribution
  - detect genes invovled in last stage.

Parameter estimation: SEM algorithm

Model

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

Parameter estimation: SEM algorithm

Model

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

1. Starting point from an heuristic.

Parameter estimation: SEM algorithm

Model

$$\begin{cases} C_i + T_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

1. Starting point from an heuristic.
2. Iteration.

$\Theta_j = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$ $\longrightarrow$ Sample $(C_i^{(j),1}, \ldots, C_i^{(j),N})$ from $\mathbb{P}_{\Theta^{(j)}}[C_i | DG_i, T_i, E_i^{\text{case}}, DE_i]$

$\downarrow$

$(\beta^{(j+1)}, \sigma^{(j+1)})$ MLE from $\mathbb{P}_{\beta,\sigma}[DG_i | DE_i, C_i^{(j)}]$

$(\kappa^{(j+1)}, \tau^{(j+1)})$ MLE from $\mathbb{P}_{\kappa,\tau}[C_i^{(j)} |, E_i^{\text{case}}, T_i]$

$\nwarrow$ (back to Θ_j)

Parameter estimation: SEM algorithm

Model

$$\begin{cases} C_i + T_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon_{i,g} \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

1. Starting point from an heuristic.
2. Iteration.

$\Theta_j = (\kappa^{(j)}, \tau^{(j)}, \beta^{(j)}, \sigma^{(j)})$ $\longrightarrow$ Sample $(C_i^{(j),1}, \ldots, C_i^{(j),N})$ from $\mathbb{P}_{\Theta^{(j)}}[C_i | DG_i, T_i, E_i^{\text{case}}, DE_i]$

$\downarrow$

$(\beta^{(j+1)}, \sigma^{(j+1)})$ MLE from $\mathbb{P}_{\beta,\sigma}[DG_i | DE_i, C_i^{(j)}]$
$(\kappa^{(j+1)}, \tau^{(j+1)})$ MLE from $\mathbb{P}_{\kappa,\tau}[C_i^{(j)} |, E_i^{\text{case}}, T_i]$

$\nwarrow$ (back to Θ_j)

3. $\hat{\Theta} = \sum_{j \geqslant \text{burn-in}} \Theta^{(j)}$.

Simulations

$$\begin{cases} C_i + T_i \sim \Gamma(k, \theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

Simulations

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

- Observed follow-up times $(T_1, \ldots, T_{150})$.
- Observed exposure $(E_1, \ldots, E_{150})$: HRT $=$ 0 or 1.

Simulations

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

- Observed follow-up times $(T_1, \ldots, T_{150})$.
- Observed exposure $(E_1, \ldots, E_{150})$: HRT $=$ 0 or 1.

- $(\tau = (2, 0.5), \kappa = (3, 0.5))$ so that:
  - Shorter last-stage for HRT=1 than HRT=0
  - 42% of positive C.
- Simulate $(LS_1, \ldots, LS_n)$ and compute $C_i = LS_i - T_i$ for each case i.

Simulations

$$\begin{cases} C_i + T_i \sim \Gamma(k,\theta) \quad \text{with} \quad k = 1 + \exp(\langle \kappa, E_i^{\text{case}} \rangle), \quad \theta = \exp(\langle \tau, E_i^{\text{case}} \rangle) \\ DG_i^g = \langle \beta^g, (1, DE_i, C_i \mathbb{I}(C_i > 0)) \rangle + \varepsilon_{i,g}, \quad \varepsilon \sim \mathcal{N}(0, \sigma_g) \end{cases}$$

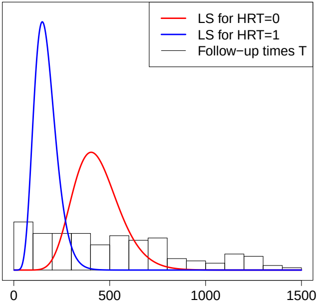
- Observed follow-up times $(T_1, \ldots, T_{150})$.
- Observed exposure $(E_1, \ldots, E_{150})$: HRT $=$ 0 or 1.

- $(\tau = (2, 0.5), \kappa = (3, 0.5))$ so that:
  - Shorter last-stage for HRT=1 than HRT=0
  - 42% of positive C.
- Simulate $(LS_1, \ldots, LS_n)$ and compute $C_i = LS_i - T_i$ for each case i.

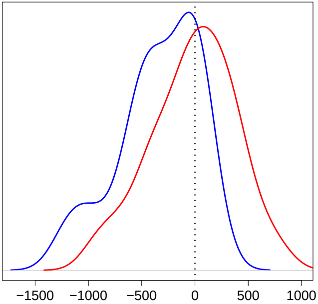
- Simulate $p = 2000$ genes. (β_0^g, β_1^g) sampled from standard gaussian distribution, (σ_g) sampled from χ^2 distribution.
- (β_2^g) sampled from $\mathcal{N}(0, 0.01)$ for $g0 = 20$ genes, and 0 for the other genes.
- Simulate DG.

Description of the simulated data (1)

Last-stage length distribution

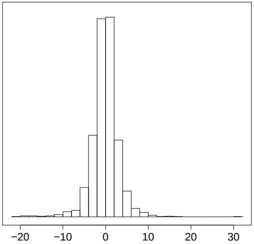


Distribution of the C_i's

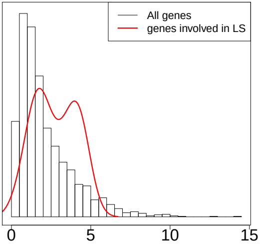


Description of the simulated data (2)

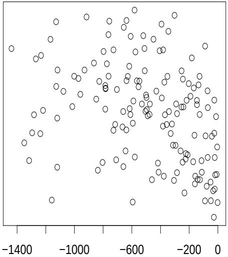
DG distribution
for one case-control pair



Gene stand. dev.
distribution

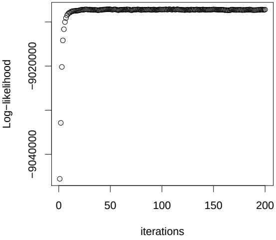


DG versus T
for one gene

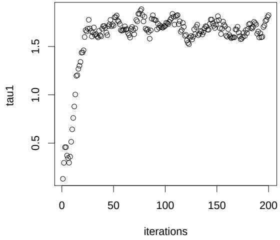


Convergence of the SEM algorithm

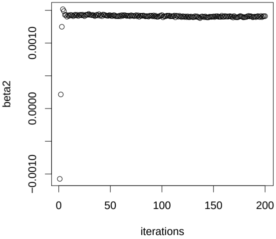
Likelihood



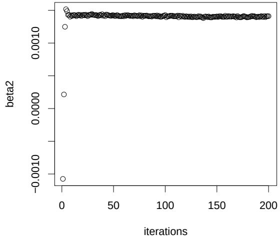
LS distribution parameter (τ_1)



β_2^g (carcinogenesis gene)



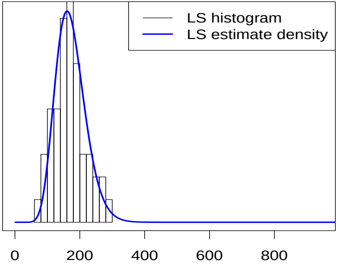
β_2^g (no carcinogenesis gene)



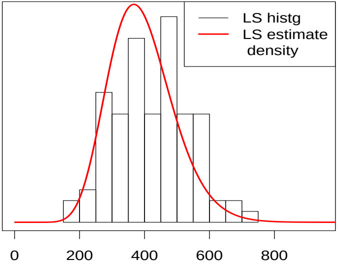
Last stage

- Histogram of the last stage LS.
- Estimated last stage density (Gamma distribution with estimated parameters): solid line.

$HRT = 1$



$HRT = 0$



Gene detection

Multiple testing

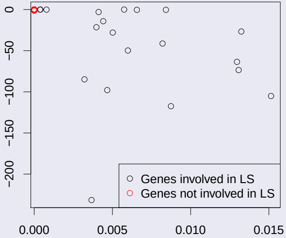
- t-test: $\beta_2^g = 0$.
- Adjust p-values (Benjamini-Hochberg).

Gene detection

Multiple testing

- t-test: $\beta_2^g = 0$.
- Adjust p-values (Benjamini-Hochberg).

$\log 10(\text{pv})$ versus β_2^g

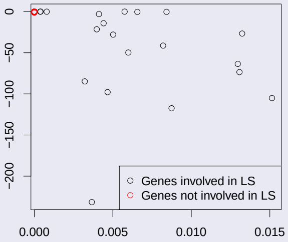


Gene detection

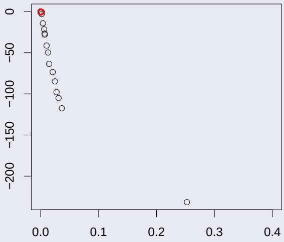
Multiple testing

- t-test: $\beta_2^g = 0$.
- Adjust p-values (Benjamini-Hochberg).

$\log 10$(pv) versus β_2^g



$\log 10$(pv) versus β_2^g/σ_g

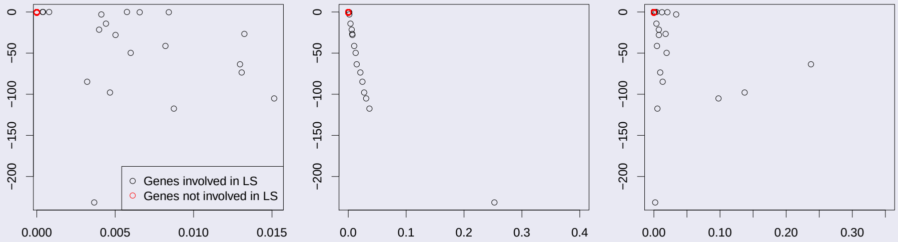


Gene detection

Multiple testing

- t-test: $\beta_2^g = 0$.
- Adjust p-values (Benjamini-Hochberg).

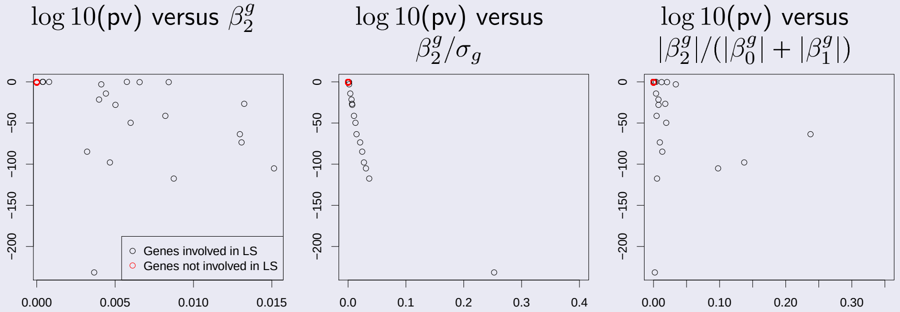
$\log 10(\text{pv})$ versus β_2^g | $\log 10(\text{pv})$ versus β_2^g/σ_g | $\log 10(\text{pv})$ versus $|\beta_2^g|/(|\beta_0^g| + |\beta_1^g|)$



Gene detection

Multiple testing

- t-test: $\beta_2^g = 0$.
- Adjust p-values (Benjamini-Hochberg).



- Detection depends on signal/noise ratio
- Detection independent on constant and exposures coefficients

Sensitivity : comparison with Spearman test

Tests

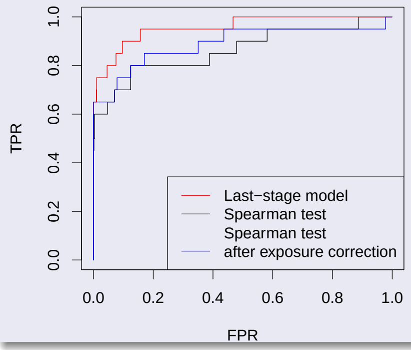
- Latent last stage model
- Spearman test (DG_g , T)
- Spearman test after correction for exposures : (Res(lm($DG_g \sim DE$)) , T)

Sensitivity : comparison with Spearman test

Tests

- Latent last stage model
- Spearman test (DG_g , T)
- Spearman test after correction for exposures : (Res(lm($DG_g \sim DE$)) , T)

ROC curve



- Higher sensitivity with latent last-stage model
- Spearman tests: higher sensitivity after correction with exposures

Further developments

- Relevant exposures

Further developments

- Relevant exposures
- Stratification
  - Type of cancer
  - Stage of cancer

Further developments

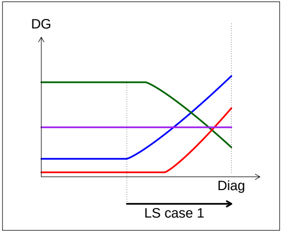
- Relevant exposures
- Stratification
  - Type of cancer
  - Stage of cancer
- Carcinogenesis driven by exposures.

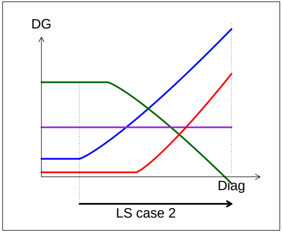
Further developments

- Relevant exposures
- Stratification
  - Type of cancer
  - Stage of cancer
- Carcinogenesis driven by exposures.
- Alternative longitudinal model.

Further developments

- Relevant exposures
- Stratification
  - Type of cancer
  - Stage of cancer
- Carcinogenesis driven by exposures.
- Alternative longitudinal model.
  $\rightarrow$ Shift in gene distribution.





Conclusion

- Statistical approach to study carcinogenesis on transcriptomics
- Flexible structure based on a linear model including a latent variable.
- Validation on simulations.
- Inclusion of biological assumptions