



Statistician's quest for biomarkers: optimizing the two stage testing procedures

Vera Djordjilović

November 22, 2019
StaTalk, Trieste

Joint work

University of Oslo

- **Magne Thoresen**
- **Jesse Hemerik**
- Christian Page
- Jon Michael Gran
- Marit Bragelien
Veierød

University of Tromsø

- Therese H. Nøst
- Torkjel M.
Sandanger



Table of Contents

Introduction

Motivating problem

ScreenMin procedure

Motivating problem revisited

Concluding remarks



Biomarkers in cancer research

- In 2018, 1 out of 6 deaths due to cancer



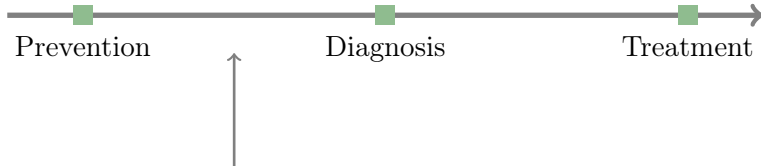
Biomarkers in cancer research

- In 2018, 1 out of 6 deaths due to cancer



Biomarkers in cancer research

- In 2018, 1 out of 6 deaths due to cancer



- Risk assessment
- Early diagnosis



Table of Contents

Introduction

Motivating problem

ScreenMin procedure

Motivating problem revisited

Concluding remarks

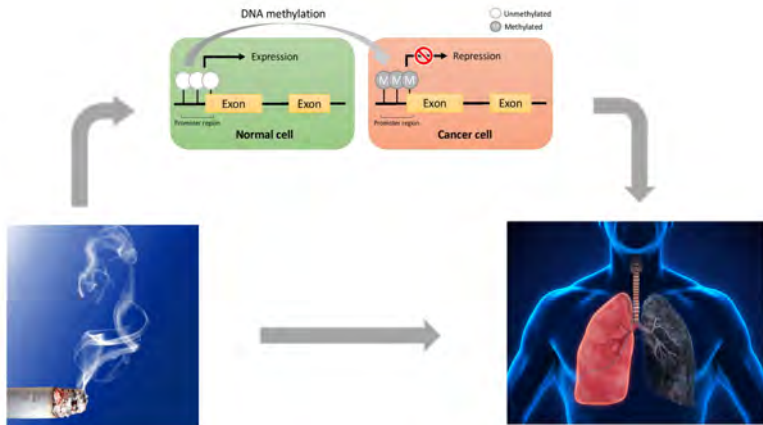


Motivating problem: lung cancer

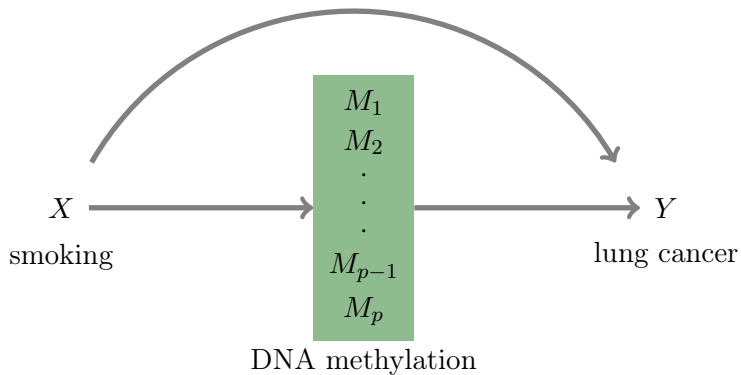
- **Lung cancer** Most common worldwide; so far no successful screening strategy.
- **Working hypothesis.** Smoking changes DNA methylation patterns, which in turn increase the risk of lung cancer.



Smoking, DNA methylation and lung cancer



The model



Mediator and the outcome model

Two building blocks:

(1) The mediator model

$$\mathbf{M}_{p \times 1} = \boldsymbol{\alpha}_0 + \boldsymbol{\alpha}X + \boldsymbol{\epsilon}_M,$$

where $\boldsymbol{\epsilon}_M \sim \mathcal{N}(0, \Sigma)$ for some positive definite matrix Σ .

(2) The outcome model

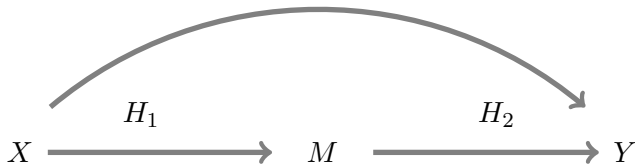
$$\text{logit} [\mathbf{P} (Y = 1)] = \beta_0 + \mathbf{M}^\top \boldsymbol{\beta} + \gamma X.$$



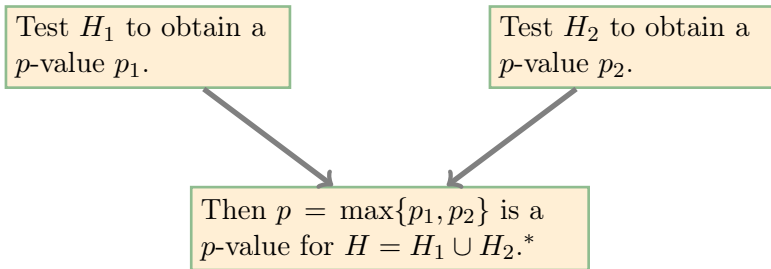
The hypothesis

To test whether M is a mediator candidate, we test H

$$H = H_1 \cup H_2.$$



The test



*Intersection union test (Gleser, 1973).



Table of Contents

Introduction

Motivating problem

ScreenMin procedure

Motivating problem revisited

Concluding remarks



Multiple potential mediators

	Test of H_{i1}	Test of H_{i2}	p -value
H_1	p_{11}	p_{12}	$\max\{p_{11}, p_{12}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
H_m	p_{m1}	p_{m2}	$\max\{p_{m1}, p_{m2}\}$



Multiple potential mediators

	Test of H_{i1}	Test of H_{i2}	p -value
H_1	p_{11}	p_{12}	$\max\{p_{11}, p_{12}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
H_m	p_{m1}	p_{m2}	$\max\{p_{m1}, p_{m2}\}$

Consider $\{\max p_i, i = 1, \dots, m\}$ and correct for multiplicity so that FWER (Bonferroni) or FDR (Benjamini and Hochberg) is controlled.



Multiple potential mediators

	Test of H_{i1}	Test of H_{i2}	p -value
H_1	p_{11}	p_{12}	$\max\{p_{11}, p_{12}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$
H_m	p_{m1}	p_{m2}	$\max\{p_{m1}, p_{m2}\}$

Consider $\{\max p_i, i = 1, \dots, m\}$ and correct for multiplicity so that FWER (Bonferroni) or FDR (Benjamini and Hochberg) is controlled.

This procedure is very conservative!



Can we do better?

Use the information on the minimum!

	Test of H_{i1}	Test of H_{i2}	$\min p$	$\max p$
H_1	p_{11}	p_{12}	$\min \{p_{11}, p_{12}\}$	$\max \{p_{11}, p_{12}\}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
H_m	p_{m1}	p_{m2}	$\min \{p_{m1}, p_{m2}\}$	$\max \{p_{m1}, p_{m2}\}$



Two step multiple testing procedure: ScreenMin

Step 1: Screening. $S = \{i : \min \{p_{i1}, p_{i2}\} < c\}$.

Step 2. Testing.

$$p_i^* = \begin{cases} |S| \max \{p_{i1}, p_{i2}\} & i \in S \\ 1 & i \notin S. \end{cases}$$



Two step multiple testing procedure: ScreenMin

Step 1: Screening. $S = \{i : \min \{p_{i1}, p_{i2}\} < c\}$.

Step 2. Testing.

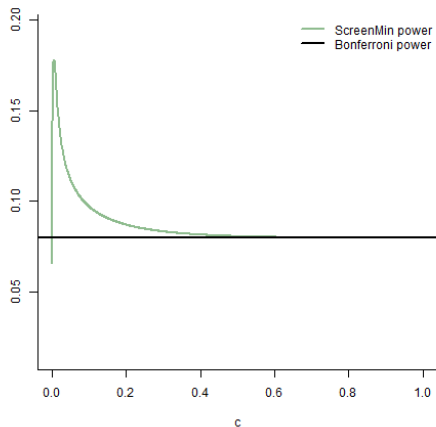
$$p_i^* = \begin{cases} |S| \max \{p_{i1}, p_{i2}\} & i \in S \\ 1 & i \notin S. \end{cases}$$

Theorem (Djordjilović et al. (2019b))

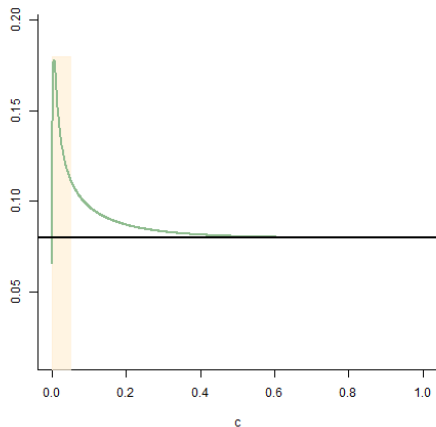
Under the assumption of independence of p -values, ScreenMin provides an asymptotic control of FWER for $\mathcal{H} = \{H_1, \dots, H_m\}$.



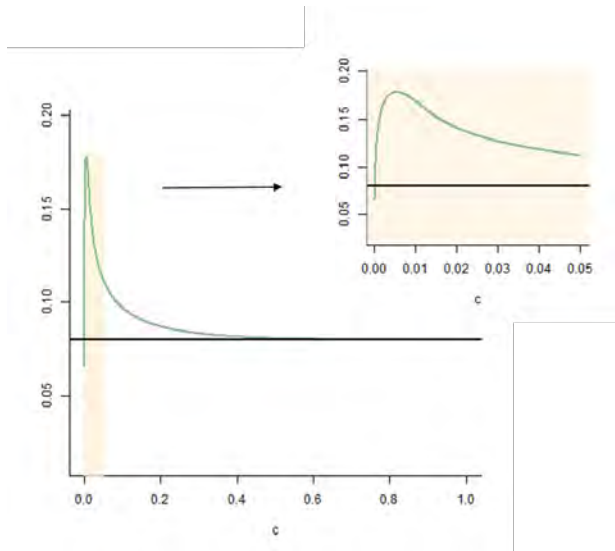
Threshold for selection c : the trade-off



Threshold for selection c : the trade-off



Threshold for selection c : the trade-off



Optimizing the threshold

For us, the optimal threshold maximizes the (average) power to reject a false hypothesis.

In general difficult, so we assume:

- Non null p -values have the same d.f. F

Then, the probability of rejection of H_i conditional on $|S|$:

$$\Pr\left(\bar{p}_i \leq \frac{\alpha}{|S|}, \underline{p}_i \leq c\right) = \begin{cases} 2F(c)F\left(\frac{\alpha}{|S|}\right) - F^2(c) & \text{for } c|S| \leq \alpha; \\ F^2\left(\frac{\alpha}{|S|}\right) & \text{for } c|S| > \alpha \end{cases}$$



Optimizing the threshold II

But not all thresholds guarantee finite sample FWER.
Constrained optimization problem:

$$\max_{0 < c \leq \alpha} \mathbb{E} \left[\Pr \left(\bar{p}_i \leq \frac{\alpha}{|S(c)|}, \underline{p}_i \leq c \right) I[|S(c)| > 0] \right]$$

subject to $\Pr(V(c) \geq 1) \leq \alpha$.



Optimizing the threshold II

But not all thresholds guarantee finite sample FWER.

Constrained optimization problem:

$$\max_{0 < c \leq \alpha} \mathbf{E} \left[\Pr \left(\bar{p}_i \leq \frac{\alpha}{|S(c)|}, \underline{p}_i \leq c \right) I[|S(c)| > 0] \right]$$

subject to $\Pr(V(c) \geq 1) \leq \alpha$.



Optimizing the threshold II

But not all thresholds guarantee finite sample FWER.
Constrained optimization problem:

$$\max_{0 < c \leq \alpha} \mathbb{E} \left[\Pr \left(\bar{p}_i \leq \frac{\alpha}{|S(c)|}, \underline{p}_i \leq c \right) I[|S(c)| > 0] \right]$$

subject to $\Pr(V(c) \geq 1) \leq \alpha$.



The (nearly) optimal threshold

No closed form solution...

However, well approximated (Djordjilović et al., 2019a) by the solution to

$$c E|S(c)| = \alpha.$$

Depends on:

- The number of considered hypotheses m ;
- Proportions of different types of hypotheses π_j , $j = 0, 1, 2$;
- Distribution of non-null p -values.



The adaptive threshold

Search for the largest $c \in (0, 1)$ such that

$$c |S(c)| \leq \alpha.$$

- Easy to compute (no numerical optimization)
- Very good approximation
- Connection with Wang et al. (2016)



Table of Contents

Introduction

Motivating problem

ScreenMin procedure

Motivating problem revisited

Concluding remarks



Smoking, DNA methylation and lung cancer

- 125 matched case-control pairs within NOWAC.
- Around 3000 CpGs, previously reported to be associated to smoking, were grouped into 72 groups, according to a gene they map to.
- Smoking coded as "Never", "Former", "Current" .
- Analysis adjusted for age, time since blood sampling, and cell composition.
- We applied the ScreenMin procedure to the 72 genes – groups of CpGs. Seven groups passed the screening.



Results

Gene	p_1	p_2
F2RL3	5.48×10^{-5}	0.54
AHRR	1.76×10^{-4}	0.57
GFI1	5.72×10^{-6}	0.42
MYO1G	6.61×10^{-6}	0.48
ITGAL	1.72×10^{-6}	0.34
VAR5	1.61×10^{-5}	0.89
CLDND1	2.37×10^{-4}	0.99

Association between smoking and methylation strong, but no evidence of association between methylation and lung cancer in the outcome model.



Table of Contents

Introduction

Motivating problem

ScreenMin procedure

Motivating problem revisited

Concluding remarks



Concluding remarks

- **Screening/selection.** In high dimensions (almost) necessary; but needs to be accounted for
- **ScreenMin.** Two stage procedure that maintains (asymptotic) FWER when testing multiple union hypotheses for arbitrary selection thresholds
- **Optimizing the threshold.** Maximizes power while guaranteeing FWER in finite samples
- **Smoking, DNA methylation and lung cancer in Norwegian women.** No evidence of mediation by DNA methylation (in blood), so no new biomarker candidates



References

- Djordjilović, V., Hemerik, J., and Thoresen, M. (2019a). Optimal two-stage testing of multiple mediators. *arXiv preprint arXiv:1911.00862*.
- Djordjilović, V., Page, C. M., Gran, J. M., Nøst, T. H., Sandanger, T. M., Veierød, M. B., and Thoresen, M. (2019b). Global test for high-dimensional mediation: Testing groups of potential mediators. *Statistics in Medicine*, 38(18):3346–3360.
- Gleser, L. (1973). On a theory of intersection union tests. *Institute of Mathematical Statistics Bulletin*, 2(233):9.
- Wang, J., Su, W., Sabatti, C., and Owen, A. B. (2016). Detecting replicating signals using adaptive filtering procedures with the application in high-throughput experiments. *arXiv preprint arXiv:1610.03330*.

