Causal Discovery of Cancer with Sparsity

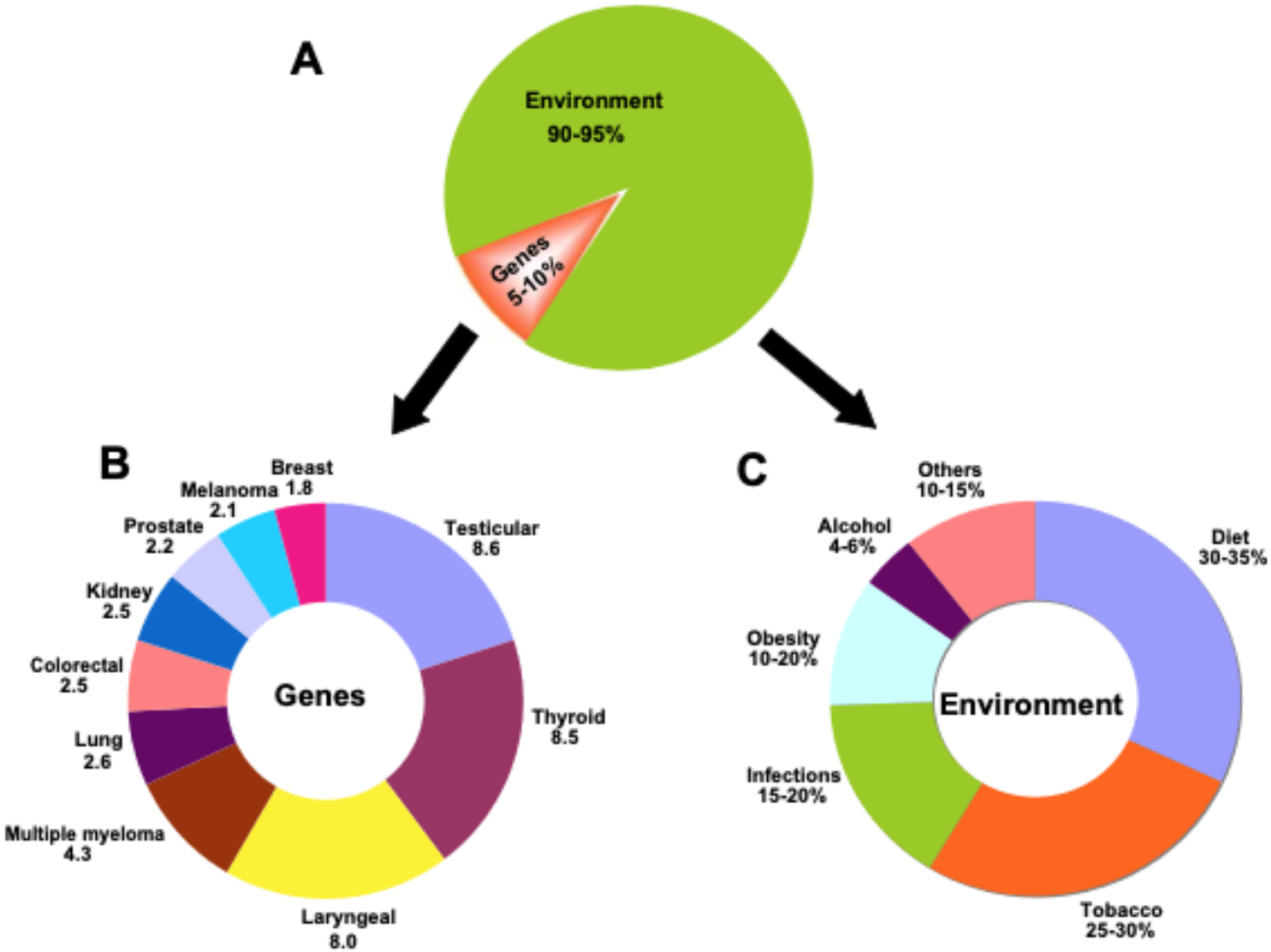
Yigit Aydede (Saint Mary's Uni.)
& **Mutlu Yuksel** (Dalhousie Uni.)
& **Andrea Giusto** (Dalhousie Uni.)
Research Portal on Machine Learning for Social and Health Policies

Applied Macro/Microeconomics Workshop - UniBZ (December 2021)

Motivation

- Even after a quarter century of extensive research, researchers are still trying to determine whether cancer is partly preventable.
- Cancer is caused by both internal factors (such as inherited mutations, hormones, and immune conditions) and environmental and lifestyle factors (such as tobacco, diet, radiation, and infectious organisms).
- Studies indicate that only 5–10% of all cancer cases are due to genetic defects and the remaining 90–95% are due to environment and lifestyle.
- For example, according to a 1997 estimate, approximately 30–40% of cancer cases worldwide were preventable by feasible dietary means (https://www.wcrf.org/diet-and-cancer/).

Split


A
Environment
90-95%
Genes
5-10%
B
Genes
Testicular
8.6
Thyroid
8.5
Laryngeal
8.0
Multiple myeloma
4.3
Lung
2.6
Colorectal
2.5
Kidney
2.5
Prostate
2.2
Melanoma
2.1
Breast
1.8
C
Environment
Diet
30-35%
Tobacco
25-30%
Infections
15-20%
Obesity
10-20%
Alcohol
4-6%
Others
10-15%

Famous quate . . .

From Craig Venter (2008) - who did the first draft sequence of the human genome

"Genes are absolutely not our fate. They can give us useful information about the increased risk of a disease, but in most cases they will not determine the actual cause of the disease, or the actual incidence of somebody getting it. Most biology will come from the complex interaction of all the proteins and cells working with environmental factors, not driven directly by the genetic code"

Evidence on "Environment"

- For each of the main cancer types, the evidence mostly comes from studies looking at a single cancer type.
- For example, the carcinogenic effects of tobacco appear to be reduced by some dietary agents. (Anand et al. 2008)
- The evidence is mixed or inconclusive for many dietary ingredients.
- For example, the link between diet and cancer is revealed by the large variation in rates of specific cancers in various countries and by the observed changes in the incidence of cancer in migrating.(https://www.wcrf.org/diet-and-cancer/)

Main goal

- Our main objective is to see whether the life-style and environmental factors can be identified as predictors of cancer with high-dimensional data.
- With a proper dataset developed to a "perfection", we want to see **if those "predictors" can be considered as "causal predictors"**.
- Judea Pearl (WHY, 2018): "ice-cream sale predicts crime rate but it's not a causal predictor"

Our contribution:

- **Developing "perfect" data for cancer research in Canada**
- **Identifying "environmental" factors with nested sparsification**

GWAS and LASSO

Genotype + *Environment* + *Genotype Environment Interactions* $\rightarrow$ *Phenotype*

Humans have:

- 3 bio base pairs of DNA
- 22,000 genes
- 661 mio SNPs (Single nucleotide polymorphisms, the most common form of genetic variation among humans),

What we do is not new in cancer research, but new for the "environment" part . . .

- **Regularization:** Variable selection via Lasso with high-dimensional proteomic data: 77 Breast cancer types with 12546 proteins
- **GWAS:** Associations between SNPs and cancer types Genome-wide association studies of cancer: current insights and future perspectives - NATURE 2018

Perfect Data?

- Survey data linked to administrative data
- No "model-leaking" across variables
- Panel in multiple time dimensions
- Unprecedented details

CCHS

- Canadian Community Health Survey (CCHS) is the largest national health survey with more than 130 thousand observations in annual files.
- It has more than 2,500 variables (in the major confidential files) that contain information ranging from the amount of weekly **carrot consumption** to the weekly time in minutes that the person spends in **Olympic weightlifting** workouts.
- But it's **useless for any type of deep statistical analysis** due to a time-confusion between features
- Infamously known as **Model-Leaking**: a naive work can discover that eating healthy strongly correlates with a chronic disease.

Unprecedented details

- Rollerblading - 12 mo
- Rollerblading - wears helmet - frequency
- Rollerblading - wears wrist protectors - frequency
- Rollerblading - wears elbow pads - frequency
- Rollerblading - wears knee pads - frequency
- Downhill skiing / snowboarding - 12 mo
- Downhill skiing - wears helmet - frequency
- Snowboarding - wears helmet - frequency
- Snowboarding - wears wrist protectors - frequency
- Skateboarding - 12 mo
- Skateboarding - wears helmet - frequency
- Skateboarding - wears wrist protectors - frequency
- Skateboarding - wears elbow pads - frequency
- Ice hockey - 12 mo
- Ice hockey - wears mouth guard - frequency
- Wears protective equipment - in-line skating - (D)
- Wears protective equipment - snowboarding - (D)
- Wears protective equipment - skateboarding - (D)
- Sun safety behaviours - Inclusion Flag - (F)
- Sunburn - 12 mo
- Time spent daily in the sun 10am to 4pm - days off - summer
- Seek shade - frequency
- Wears hat - frequency
- Wears long pants / skirt - frequency
- Wears sunglasses - frequency
- Uses sunscreen on face - frequency
- Uses sunscreen on face - SPF
- Uses sunscreen on body - frequency
- Uses sunscreen on body - SPF
- Used tanning bed or booth - 12 mo
- Used tanning bed or booth - frequency
- Used tanning bed or booth - reporting period
- Used tanning bed or booth - main reason
- Respondent protects self from sun - (D)

- Smoking - Inclusion Flag - (F)
- Type of smoker (daily / occasionally / not at all) - presently
- Smoked - 30 d
- Smoked daily - 30 d
- Smoked more than 100 cigarettes - lifetime
- Smoked a whole cigarette - lifetime
- Smoked daily - lifetime (occasional / former smoker)
- Age - smoked first whole cigarette
- Age - began smoking daily (daily / former daily smoker)
- Num of cigarettes smoked daily (daily smoker)
- Num of cigarettes smoked daily (occasional smoker)
- Num of days - smoked 1 cigarette or more (occasional smoker) - 1 mo
- Stopped smoking - when (former occasional smoker)
- Stopped smoking - month (former occasional smoker)
- Stopped smoking - num of years (former occasional smoker)
- Num of cigarettes smoked daily (former daily smoker)
- Stopped smoking daily - when (former daily smoker)
- Stopped smoking daily - month (former daily smoker)
- Stopped smoking daily - num of years (former daily smoker)
- Quit smoking completely (former daily smoker)
- Quit smoking completely - when (former daily smoker)
- Quit smoking completely - month (former daily smoker)
- Quit smoking completely - num of years (former daily smoker)
- Smoking status (type 2) - traditional definition - (D)
- Num of years respondent has smoked daily - (D)
- Num of years since stopped smoking completely - (D)
- Smoking cessation methods - Inclusion Flag - (F)
- Reduced / quit smoking - nicotine patch - 12 mo
- Usefulness of nicotine patch
- Reduced / quit smoking - nicotine gum - 12 mo
- Usefulness of nicotine gum
- Reduced / quit smoking - medication - 12 mo
- Usefulness of medication
- Stopped smoking for at least 24 hours - 12 mo
- Tried to reduce / quit smoking - nicotine patch - 12 mo
- Tried to reduce / quit smoking - nicotine gum - 12 mo
- Tried to reduce / quit smoking - medication - 12 mo
- Attempted / successful quitting smoking - (D)
- Tobacco products alternatives - Inclusion Flag - (F)
- Smoked little cigars / cigarillos - 30 d
- Smoked little cigars / cigarillos - plain / flavoured
- Smoked other cigars - 30 d
- Used an electronic cigarette - 30 d
- Smoked a pipe - 30 d
- Used chewing tobacco / pinch / snuff - 30 d
- Smoked a tobacco water-pipe - 30 d
- Alternative tobacco product usage - (D)
- Exposure to second-hand smoke - Inclusion Flag - (F)
- Someone smokes inside home every day
- Number of people who smoke inside home
- Smoking allowed inside home
- Smoking restrictions inside home
- Smoking restrictions - allowed in certain rooms only
- Smoking restrictions - restricted in the presence of young children
- Smoking restrictions - allowed if windows are open
- Smoking restrictions - other

Discharge Abstract Database (DAD)

- Developed in 1963, DAD captures administrative, clinical and demographic information on hospital discharges (including deaths, sign-outs and transfers, day surgeries).
- Data are received from all acute care facilities across Canada (except Quebec)
- Selected chronic care, rehabilitation, and psychiatric facilities also submit data to DAD.
- Currently, more than 3 million records are submitted to DAD annually.
- **Length of Stay, Patient Demographics, Admission Data, Discharge Data, Patient Service Information, Service Transfers, Provider Information, Diagnosis Information, Intervention Information, Special Care Information, Blood Information, Reproductive Care Information**

Linking CCHS to DAD

- First time in Canada!
- We can follow the same person in DAD for years
- The details in diagnostic information is captured in more than 33,000 codes with timing in days and hours
- We linked DAD and each CCHS so that we can trace each person's DAD in each CCHS between 2002 and 2011.
- For example, 2001 CCHS linked to DAD has the medical history of each of 130000 people between 2002 and 2011
- We linked each CCHS between 2001 and 2008 to DAD

"Perfection" in data

Example:

- We dropped people who were cancer patients in 2001 CCHS or had gone through a cancer treatment prior to 2001.
- Among more than 60 thousand people who have no cancer in 2001 and never had a cancer before 2001, we identified around 6 thousand people who had developed a cancer in the following 9 years.

With the same process:

- We linked CCHS from 2001 to 2008 to DAD and pooled them in one dataset.
- In the pooled data, we have more than **500K people about 40,000 of whom became cancer patients** later in their life.
- We have each cancer patients' full medical history (cancer related or not) before and after the cancer diagnosis.
- About 5% of cancer patients have multiple cancer types at Stage 1.

Available statistical tools

- With this data set, we have an opportunity to use every feature in the survey without having a concern about possible reverse causality problems or model-leaking issues.
- This dataset can be used any chronic disease - not only cancer research
- But, one can have more than 200 thousand features in a parametric model by including only first-level selective interactions and first-degree polynomials.
- We need a dimension reduction for parametric models but their linearity (model design) should be verified with nonparametric predictive models.

Regularized parametric models

There are multiple objectives in regularized parametric models:

- Better prediction by preventing over-fitting,
- Dimension reduction
- Identifying the true sparsity in the model (subset of variables related to outcome).

These are not mutually exclusive, but most of the economics literature relates to the last two objectives.

- The first objective may seem questionable, as the base function parametric and linear.
- Although penalized linear regressions are in the bottom of ranking in accuracy, they are the best in interpretability.

lasso (least absolute shrinkage and selection operator)

The lasso coefficients minimize the following quantity:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

- The ℓ_1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large.
- Hence, unlike Ridge, the lasso performs variable selection.

Sparsity

The lasso estimators have a special property: their solutions are sparse, i.e., at a solution $\widehat{\beta}$ we will have $\widehat{\beta}_j = 0$ for many components $j \in \{1, \ldots, p\}$.

- Note that sparsity is desirable for two reasons:
  - **(i)** it corresponds to performing variable selection in the constructed linear model, and
  - **(ii)** it provides a level of interpretability (beyond sheer accuracy)
- Lasso is known to identify the true sparsity when the underlying model is approximately sparse.
- With some additional technical assumptions, the lasso estimator is **"sparsistent"** (Charpentier, 2019).

Lasso - Oracle Estimators

- Two objectives in using penalized regressions:
  (i) model selection or **identifying the "correct" sparsity**,
  (ii) and pure prediction or **forecasting accuracy**.
- In the model selection, the objective is to shrink the dimension of the model to the "true" sparsity. This is usually evaluated by checking whether the **Oracle properties** are satisfied.

Oracle properties:

- Ideally, we would like to have a procedure which classifies all truly zero coefficients EXACTLY as zero with probability tending to one.
- ... and whose asymptotic distribution for the truly non-zero coefficients is the same as if only the variables pertaining to these had been included from the outset.

The literature tells that Lasso is not an "Oracle" estimator.
Adaptive Lasso was developed (Zou 2006) to fill this gap.

Adaptive Lasso

$$L(\beta) = \sum_{i=1}^{n} \left(y_i - x_i'\beta\right)^2 + \lambda_n \sum_{j=1}^{p} \frac{1}{w_j} |\beta_j|$$

- The weights are more "**intelligent**" than those for the plain Lasso.
- The plain Lasso penalizes all parameters equally . . . while the adaptive Lasso is likely to penalize non-zero coefficients less than the zero ones.

Downside: Two-step procedure as opposed to the one-step plain Lasso. And, OLS may not be a good initial estimator

See Hui Zou (2006), The Adaptive Lasso and Its Oracle Properties

Thresholded LASSO

The thresholded Lasso estimator, $\tilde{\beta}_j$ is defined as follows

$$\tilde{\beta}_j = \hat{\beta}_j \qquad \text{if } \left|\hat{\beta}_j\right| \geq T$$
$$\tilde{\beta}_j = 0 \qquad \text{if } \left|\hat{\beta}_j\right| < T$$

where T could be defined as a grid, for example, $T \in \{0.01, 0.1, 1, 2, 10\}$. And with a grid search on λ and T that minimizes BIC:

$$BIC_{\lambda T} = \log\left(SSE_{\lambda T}\right) + \frac{\log(n)}{n} nz_{\lambda T}$$

$nz_{\lambda T}$ is the number of non-zero coefficients.

Callot, Caner and Kock (2015) show that thresholded Lasso yields substantial improvements in terms of variable selection.

Group Lasso

We can see if a "group" of variables are sparsified or not with adaptive group lasso

The adaptive group lasso penalty is defined as

$$\Phi(\beta) = (1-\lambda)\sum_{J=1}^{m}\gamma_J\left\|\beta^{(J)}\right\|_2 + \lambda\sum_{i=1}^{p}\xi_i\left|\beta_i\right|$$

where $\lambda \in [0,1], \gamma \in [0,\infty)^m$ are the group weights, and parameter weights $\xi = \left(\xi^{(1)},\ldots,\xi^{(m)}\right) \in [0,\infty)^p$ for $\xi^{(1)} \in [0,\infty)^{p_1},\ldots,\xi^{(m)} \in [0,\infty)^{p_m}$. As with the elastic net method, the tuning parameter α could lead to two different methods by taking $\lambda = 1$ (lasso penalty) or $\lambda = 0$ (group lasso penalty).

Codes used by Cancer Registrars ICD-0-3

https://training.seer.cancer.gov/icd10cm/appendix-b/

- Lip, Oral Cavity and Pharynx (C00-C14)
- Digestive Organs (C15-C26)
- Respiratory System and Intrathoracic Organs (C30-C39)
- Coding for Bones, Joints and Articular Cartilage (C40-C41)
- Skin (Melanoma, Merkel Cell, and Other Skin Histologies) (C43, C44, C4a)
- Kaposi Sarcoma (9140)
- Peripheral Nerves, Retroperitoneum, Peritoneum, and Soft Tissues (C47, C48, C49)
- Breast and Female Genital Organs (C50 – C58)
- Male Genital Organs (C60-C63)
- Urinary Tract (C64-C68)
- Eye, Brain and Other Parts of the Central Nervous System (C69-C72)
- Thyroid, Other Endocrine Glands, and Ill-defined Sites (C73-C76)
- Lymph Nodes, Secondary Cancers & Unknown Primary Site (C77-80)
- Reportable Benign (/0 or /1) Neoplasms

Respiratory System and Intrathoracic Organs (C30-C39)

ICD-O-3 Code	ICD-O-3 Description	ICD-10-CM	ICD-10	ICD-9-CM
C30_	Nasal Cavity and Middle Ear	C30.	C30.-	160.-
C31_	Accessory Sinuses	C31.	C31.-	160.-
C32_	Larynx	C32.	C32.-	161.0
C339	Trachea	C33	C33	162
C34_	Lung	C34.	C34.-	162.-
C379	Thymus	C37	C37	164.0
C38_	Heart, Mediastinum	C38.-	C38.-	164.-
C39_	Other and ill-defined sites within respiratory system and intrathoracic organs	C39.-	C39.-	165.-

Our application

- We could also do multinomial lasso but we wanted to customize the nesting structure
- We first define 3 layers of nested models identified by a different binary outcome
- For "cancer", y^{c0}, one for each 14 cancer groups(g), y^{cg} and each type (t), y_t^{cg} where $t = \{1, \cdots, 118\}$
- We set $1(c) + 14(g) + 118(t)$ models using the same set of variables in each
- Hence, the only difference is the outcome variable across $1(c) + 14(g) + 118(t)$ models
- We also have several other alternative nesting structures, like only breast cancers vs. all or only sub-types of cancers related to respiratory system,
- We apply variable selection with Thresholded (group) Lasso and identify $1(c) + 14(g) + 118(t)$ **nested** sparsified models

Nested & Thresholded Lasso

$$y^{c0} = \cdots + \beta_2 x_2 + \cdots + \beta_{75} x_{75} + \cdots + \beta_{98} x_{98} + \cdots$$
$$y^{c1} = \cdots + \beta_2 x_2 + \cdots + \beta_5 x_5 + \cdots + \beta_{901} x_{901} + \cdots$$
$$\vdots$$
$$y^{c14} = \cdots + \beta_2 x_2 + \cdots + \beta_{51} x_{51} + \cdots + \beta_{7015} x_{7015} + \cdots$$
$$y_1^{c1} = \cdots + \beta_2 x_2 + \cdots + \beta_{16} x_{16} + \cdots + \beta_{9521} x_{9521} + \cdots$$
$$\vdots$$
$$y_{118}^{c14} = \cdots + \beta_2 x_2 + \cdots + \beta_{26} x_{26} + \cdots + \beta_{1809} x_{1809} + \cdots$$

Process?

- We identify overlapping and nonoverlapping variables with non-zero coefficients.
- $X_{12561} \in \{y_{12}, y_{29}, y_{101}\}$ Cancer types 12, 29, and 101 share the factor X_{12561}
- We also find common factors for "cancer" by y^c and 14 main categories.
- This leads to a big approximately $200,000 \times 133$ matrix where columns are types (nested) and the rows are features
- Non-zero elements by rows representing what factors are mutually common in $1(c) + 14(g) + 118(t)$ types
- The differences in coefficients (normalized) represent relative importance of each factor
- Age is the main predictor for all types
- Since more than 80% of people who had cancer (any type) in following years are between 55 and 75, we restrict the data only for that age group.

Cons and Pros

Pros:

- The same data: effects of data imperfections should be reduced
- Same models: effects of model imperfections should be reduced
- Thus, results reveal common and unique factors across cancer types better than individual studies
- The data provides a better quality (size, richness, panel dimension) than controlled experiments

Cons:

- Models are LPM. Nonlinearity with this size data cannot be captured by Lasso
- Nonparamteric models (AdaBoost) have a better prediction accuracy indicating nonlinearities
- Computation time for each lasso estimation (x133) is long.
- Data is confidential and can be accessed only in Research Data Centers: no super computers!

Predictive power - Cancer as a "common" disease

Any algorithm with AUC larger than 0.70 can be used as a **"Recommender System"** in medical practice

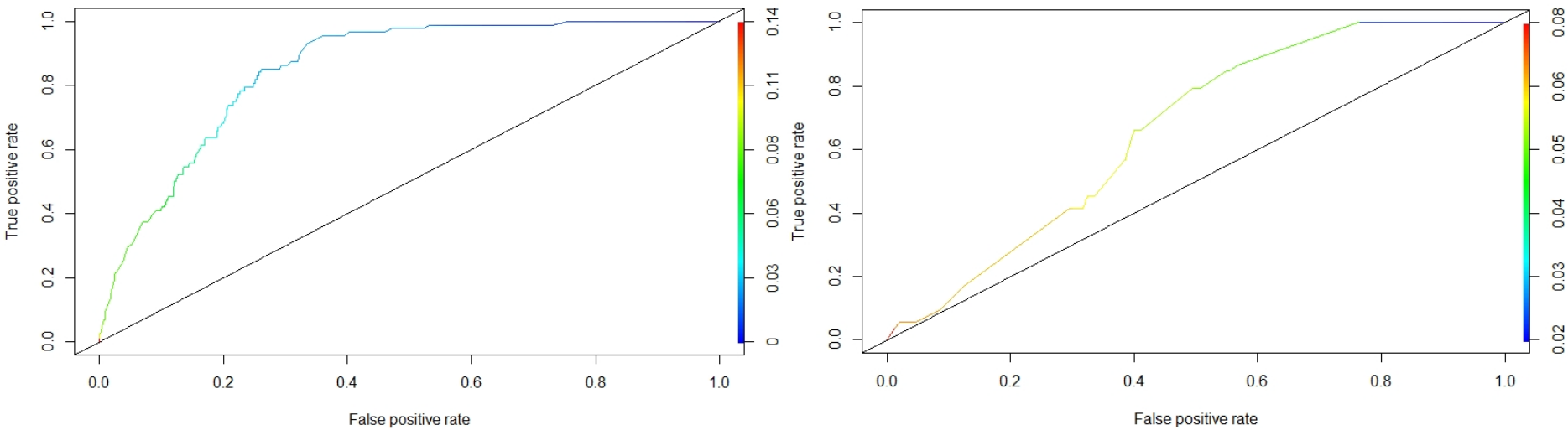
With Age	Cancer	Doesn't
pred>th	0.9318	0.3355
pred<th	0.0682	0.6645

Optimal Threshold
0.0238971
AUC
0.833201

55-75	Cancer	Doesn't
pred>th	0.8491	0.5487
pred>th	0.1509	0.4513

Optimal Threshold
0.05185311
AUC
0.7314851

ROC's tell a less optimistic story



AdaBoost improves the results (AUC) at least 10 PP indicating a strong nonlinearity

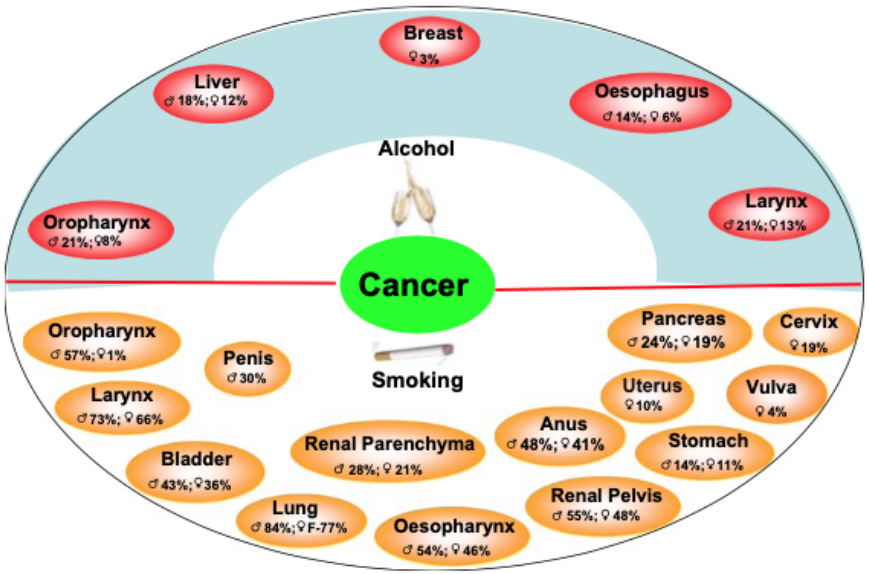
Causal discovery

Our findings verify some of the most recommended factors with substantial details.

Selected findings:

- Alcohol consumption is a factor for C15-C26 (digestive organs) but limited to excessive drinking for a long period
- Being active reduces all cancer types: Cardio-type workouts less effective than those related to strength training. Less than 120 min activity in a week has no effect
- BMI is a factor but depending on age, duration, and level
- Vegetable consumption is good, specially tomato and carrot

Smoking . . .



Cigarette smoking is listed as the major single preventable cause of cancer in the United States, estimating that cigarette smoking accounted for about 30% of **all cancer deaths**. Doll and Peto

We found:

- **Smoking is not a common factor except for very few types of cancer**
- Second-hand smoking (SHS) is as bad as the first-hand smoking (FHS)
- Smoking in the past has marks in the future that is not repairable.
- Even exposure to a lower-degree SHS has strong effects
- After certain time, quieting smoking has little impact

. . .

Contradictory Findings

- After controlling for age (55-75), there are few predictors selected by sparsified models,
- When we include "age" as a predictor, we have even more sparsification (3x)
- Smoking may not be a common factor for most cancer types.
- The consumption of read meat is not selected in any sparsified model,
- Stress and sleep deficiency are strong predictors
- A very few healthy eating habits are associated with cancer ...

Concluding remarks

- We have a long way to go ...
- We are helped with cancer researchers about our data and findings
- Nested spasifications layered with outcome differentials can be used for model selection
- Differences in nested and non-nested models may help causal discoveries
- We work on hazard models to see if the duration until the cancer diagnosis can be predicted