XAI in Epidemiology - Exploring Interacting Causes of a Health Outcome with **CoOL** (**C**auses **o**f **O**utcome **L**earning)

Andreas Rieckmann[1], Piotr Dworzynski[1], Leila Arras[2], Sebastian Lapuschkin[2], Wojciech Samek[2], Onyebuchi A. Arah[3], Naja H. Rod[1], Claus T. Ekstrøm[4]

[1]Section of Epidemiology, Department of Public Health & Novo Nordisk Foundation, University of Copenhagen, Denmark
[2]Department of Artificial Intelligence, Fraunhoher Heinrich Hertz Institute & BIFOLD, Berlin, Germany
[3]Department of Epidemiology, University of California & Department of Statistics, UCLA College, USA
[4]Section of Biostatistics, Department of Public Health, University of Copenhagen, Denmark

TASK — Discover (interacting) exposures causing a (health) outcome

Tabular data:

Binary Exposures: X_i

- sex (male/female)
- age category
- taking drug A
- taking drug B
- physically active
- smoking
- high BMI
- high blood pressure
- exposed to pollutant C
- ...

Questions:

- **Which exposures** cause an increased risk?
- Do they act **alone or in synergy**?

Health Outcome: Y

Disease (1=Yes or 0=No)

Note: a similar task might also occur in other domains (social sciences, environment, public policy, finance, marketing, industry...)

Synergy/Interaction example:

$P(Y=1\|X_1,X_2)$		$X_1=0$ No Asbestos $\bar{A}$	$X_1=1$ Asbestos A
$X_2=0$	No Smoking $\bar{S}$	0.11 %	0.67 %
$X_2=1$	Smoking S	0.95 %	4.50 %

Lung cancer risk

baseline risk = $P(Y=1|X_1=0, X_2=0)$

Interaction Coefficient:

$$IC = (R_{A\bar{S}} - R_{\bar{A}\bar{S}}) - [(R_{A\bar{S}} - R_{\bar{A}\bar{S}}) + (R_{\bar{A}S} - R_{\bar{A}\bar{S}})]$$
$$= R_{AS} - R_{A\bar{S}} - R_{\bar{A}S} + R_{\bar{A}\bar{S}}$$
$$= 4.5 - 0.67 - 0.95 + 0.11 = 2.99$$

⇒ if IC > 0 positive interaction
if IC < 0 negative interaction
if IC = 0 no interaction

Can we discover such interactions with machine learning?

STANDARD SOLUTION — "Linear" regression

Model:

$$P(Y|X_1=x_1, X_2=x_2) = c_0 + c_1 \cdot x_1 + c_2 \cdot x_2 + c_3 \cdot x_1 x_2$$

regression terms = all possible combinations of input variables!

$c_0 = P(Y|X_1=0, X_2=0)$ = baseline risk

$c_1 = P(Y|X_1=1, X_2=0) - P(Y|X_1=0, X_2=0)$ = risk due to X_1 alone

$c_2 = P(Y|X_1=0, X_2=1) - P(Y|X_1=0, X_2=0)$ = risk due to X_2 alone

$c_3 = P(Y|X_1=1, X_2=1) - P(Y|X_1=1, X_2=0) - P(Y|X_1=0, X_2=1) + P(Y|X_1=0, X_2=0)$

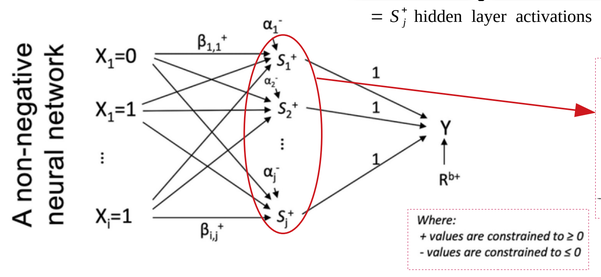
= risk due to interaction between X_1 and X_2 = IC

Limitations:
- 2^N regression terms (11 binary exposures → 2048 terms!)
- overfitting
- requires large dataset
- computationally challenging
- hard to interpret variables occurring in multiple interaction terms
- misleading results even if p-value of regression coefficients is low

OUR APPROACH — Neural network + XAI

$$P(Y=1|X_1, X_2, ...) = \sum_j \underbrace{\left(\text{ReLU}\left(\sum_i X_i \cdot \beta^+_{i,j} + \alpha^-_j \right) \right)}_{= S^+_j \text{ hidden layer activations}} + R^{b+}$$

R^{b+} = baseline risk

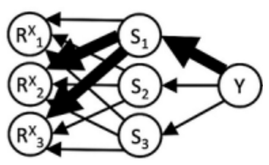


Non-linear hidden layer can modelize any higher-order interaction (no more need to test all terms explicitly!), as well as standalone exposure effects

- Stochastic Gradient Descent
- Squared prediction-error loss $(Y_{true} - \hat{P}_{Model}(Y|\boldsymbol{X}))^2$
- Regularization through squared L2-norm penalty on weights $\|\boldsymbol{\beta}\|_2^2$
- Initialize baseline risk parameter with mean risk $E[Y_{true}]$

METHOD — Discover high-risk subgroups

Step 1: **Decompose the prediction** for each individual into risk contributions per exposure (using Layer-wise Relevance Propagation) **[Bach et al. 2015]**



R_i R_j R_{total}

- Output layer: $R_{total} = P_{Model}(Y=1|\boldsymbol{X}) - R^{b+}$ (baseline risk)
- Hidden layer: $R_j = \frac{S_j}{\sum_{j'} S_{j'}} R_{total}$
- Input layer: $R_i = \sum_j \frac{X_i \cdot \beta^+_{i,j}}{\sum_{i'} X_{i'} \cdot \beta^+_{i',j}} R_j$

Step 2: **Cluster individuals** based on their risk contributions (using Manhattan distances and the Ward's method for hierarchical clustering) **[Strauss et al. 2017]**

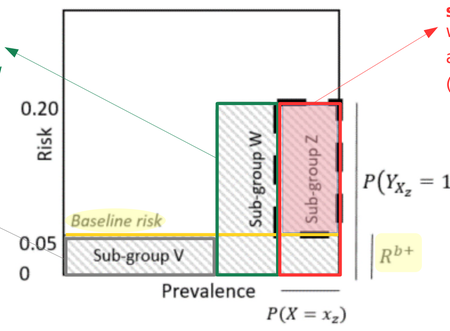
Id	R^{b+}	R^{Man}	R^{Woman}	$R^{Drug\ A}$	$R^{No\ drug\ A}$	$R^{Drug\ B}$	$R^{No\ drug\ B}$	
1	0.05	0	0	0	0	0	0	subgroup
4	0.05	0	0	0	0	0	0	
...	...	...	...	...	...	...	...	
2	0.05	0.075	0	0.075	0	0	0	subgroup
5	0.05	0.075	0	0.075	0	0	0	
...	...	...	...	...	...	...	...	
3	0.05	0	0.075	0	0	0.075	0	subgroup
6	0.05	0	0.075	0	0	0.075	0	
...	...	...	...	...	...	...	...	

Result 1: Plot of **prevalence and mean-risk** per subgroup

subgroup 3
with mean-risk > baseline risk
and a corresponding set of exposures X_W
(typically low prevalence)

subgroup 1
with mean-risk = baseline risk
and no exposure X_V=0
(typically large prevalence)

subgroup 2
with mean-risk > baseline risk
and a corresponding set of exposures X_Z
(typically low prevalence)



Excess fraction for subgroup 2:

$$\frac{P(\boldsymbol{X}=\boldsymbol{x}_z) \cdot (P(Y=1|\boldsymbol{X}=\boldsymbol{x}_z) - R^{b+})}{P(Y=1)} =$$

= "public health impact"

whole area = = mean-risk in population = $P(Y=1) = P(Y=1, \boldsymbol{X}=\boldsymbol{0}) + P(Y=1, \boldsymbol{X}=\boldsymbol{x}_w) + P(Y=1, \boldsymbol{X}=\boldsymbol{x}_z)$

Result 2: Table of **mean-risk contributions per exposure** for each subgroup

Mean risk contributions by sub-group (Standard deviation) [mean risk contribution if other exposures are set to 0]	Baseline_risk	sex_0	sex_1	drug_a_0	drug_a_1	drug_b_0	drug_b_1
subgroup 1	4.6% (0.0%) [4.6%]						
subgroup 2	4.6% (0.0%) [4.6%]		7.7% (0.0%) [0.0%]				7.7% (0.0%) [0.0%]
subgroup 3	4.6% (0.0%) [4.6%]	8% (0.1%) [0.0%]			7.7% (0.0%) [0.0%]		

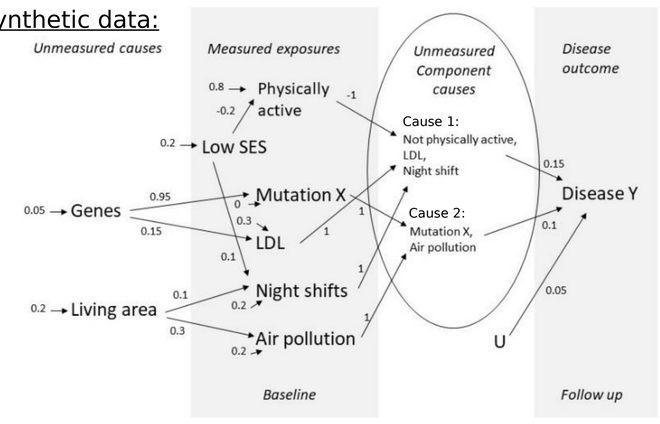
How to test exposures for synergy?

if **"contribution of exposure X_i within subgroup" > "contribution of exposure X_i with other exposures set to 0"**

⇒ then **synergy of X_i with other exposures in the subgroup**

EXPERIMENTS — Simulation

Synthetic data:



Ground truth risks:

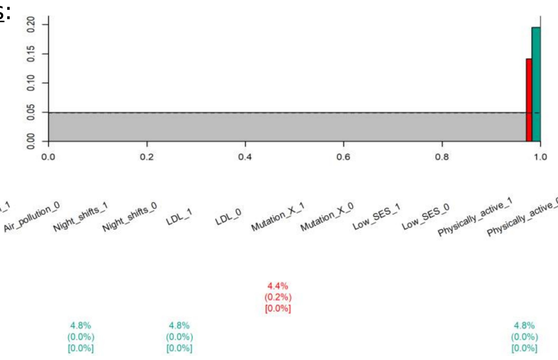
P(Y) = 5,4 % = mean-risk in population

P(Y | U) = 5 % = baseline risk

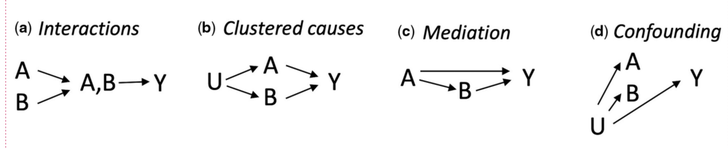
P(Cause 1) = 1,8 % = prevalence of subgroup with Cause 1
P(Y | Cause 1) = 15 % = mean-risk for this subgroup

P(Cause 2) = 1,2 % = prevalence of subgroup with Cause 2
P(Y | Cause 2) = 10 % = mean-risk for this subgroup

CoOL Results:



More simulations and robustness checks are available in the Supplementary Material of the paper with below causal structures:



Publication:

Rieckmann et al. Causes of Outcome Learning: a causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome. International Journal of Epidemiology, 2022

- Code: https://cran.r-project.org/package=CoOL
- Website: https://www.causesofoutcomelearning.org