medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Causes of Outcome Learning:

A causal inference-inspired machine learning approach to disentangling common combinations of potential causes of a health outcome

A Rieckmann[1], P Dworzynski[2], L Arras[3], S Lapuschkin[3], W Samek[3], OA Arah[4], NH Rod[1], CT Ekstrøm[5]

[1] Section of Epidemiology, Department of Public Health, University of Copenhagen

[2] Novo Nordisk Foundation Center for Basic Metabolic Research, University of Copenhagen

[3] Machine Learning Group, Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute

[4] Department of Epidemiology, Fielding School of Public Health, University of California, Los Angeles and Department of Statistics, UCLA College of Letters and Science, Los Angeles

[5] Section of Biostatistics, Department of Public Health, University of Copenhagen

Abstract: 248 words.
Manuscript: 4925 words.
Figures: 5
Supplementary table: 1, supplementary information: 4, supplementary simulation: 9, Supplementary real life data analysis: 1.

Key words: Causes of effects, Sufficient component cause model, inductive-deductive, Machine Learning, Neural networks, Explanations, Precision Public Health, Complex epidemiology, Interactions, Syndemics, Supervised clustering

Conflict of interest: None.

Funding: AR was supported by an international postdoc grant by the Independent Research Fund Denmark (9034-00006B). PD was supported by a research grant from the Danish Diabetes Academy funded by the Novo Nordisk Foundation. LA, SL, and WS are supported by the German Ministry for Education and Research as BIFOLD (refs. 01IS18025A and 01IS18037A) and TraMeExCo (ref. 01IS18056A). OAA was supported by a grant (R01EB0276502) from the National Institute of Biomedical Imaging and Bioengineering (NIBIB), a grant (UL1TR001881) from the National Center for Advancing Translational Sciences (NCATS), both at the National Institutes of Health (NIH).

Acknowledgment: The authors would like to thank the colleagues at Section of Epidemiology, Department of Public Health, University of Copenhagen for valuable comments and suggestions on the idea throughout the development. The authors are grateful to Rasmus Wibæk Christensen, Stine Byberg and Douglas Ezra Morrison for comments on the manuscript, to Tue Kjærgaard Nielsen for assistance with the implementation of dendrograms, and to Thorkild IA Sørensen for guidance on literature regarding body mass index, blood pressure and mortality.

Pre-print not peer reviewed nor accepted by a journal)

NOTE: This preprint reports new research that has not been certified by peer review and should not be used to guide clinical practice.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Abstract

Nearly all diseases can be caused by different combinations of exposures. Yet, most epidemiological studies focus on the causal effect of a single exposure on an outcome. We present the Causes of Outcome Learning (CoOL) approach, which seeks to identify combinations of exposures (which can be interpreted causally if all causal assumptions are met) that could be responsible for an increased risk of a health outcome in population sub-groups. The approach allows for exposures acting alone and in synergy with others. It involves (a) a pre-computational phase that proposes a causal model; (b) a computational phase with three steps, namely (i) analytically fitting a non-negative additive model, (ii) decomposing risk contributions, and (iii) clustering individuals based on the risk contributions into sub-groups based on the predefined causal model; and (c) a post-computational phase on hypothesis development and validation by triangulation on new data before eventually updating the causal model. The computational phase uses a tailored neural network for the non-negative additive model and Layer-wise Relevance Propagation for the risk decomposition through this model. We demonstrate the approach on simulated and real-life data using the R package 'CoOL'. The presentation is focused on binary exposures and outcomes but can be extended to other measurement types. This approach encourages and enables epidemiologists to identify combinations of pre-outcome exposures as potential causes of the health outcome of interest. Expanding our ability to discover complex causes could eventually result in more effective, targeted, and informed interventions prioritized for their public health impact.

Introduction

Many diseases may be caused by several different combinations of exposures. As putative causes, such exposures may act together and lead to a combined effect that exceeds the sum of the individual effects, also called synergism. A common example of synergism is how the combined effect of smoking and asbestos on lung cancer exceeds the sum of their individual effects.[1] The most established theoretical framework for studying synergistic effects of multiple causes in epidemiology is the *sufficient cause model,*[2]. Assessment of synergism between causes may provide etiological insight into how to prevent and treat disease. It may also help to identify and quantify the disease burden in high-risk sub-groups. Thus, understanding the spectra of exposures rather than single exposures for effective preventive strategies has been highlighted as essential for decades. Rose for example says "*… risk assessment must consider all relevant factors together rather than confine attention to a single test, for nearly all diseases are multifactorial*" when discussing effective policy decisions.[3]

Despite the policy relevance, few epidemiological studies have analytically tried to identify combination of causes for specific outcomes. We suspect that the apparent lack of epidemiological studies for questions about causes of outcomes is due to frequently taught frameworks for epidemiologists that warn against type 1 errors from multiple testing (false positive findings),[4] various confounding structures for each exposure,[5] the overwhelming number of combinations between exposures that can be created,[6] and lack of established theoretically founded approaches for applied data analysis,[6,7] though some do exist.[8–10]

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

We introduce a machine learning- and causal inference-based approach called the Causes of Outcome Learning (CoOL) approach that attempts to generate insights into questions like *"Given a particular health outcome, what are the most common combinations of exposures, which might have been its causes?"*. We present the approach assisted by a simple simulated example. A step-by-step tutorial is included in Supplementary simulations 1-6, robustness checks in Supplementary simulations 7-9, and a real-life application in Supplementary real life data analysis. Terms that may need additional explanation are marked with * and explained in the Supplementary Table 1.

Motivating example

We use a simple simulation setting to motivate the CoOL approach. We generate a study population of 10,000 individuals, of whom 50% are men and 50% are women, 20% are exposed to drug A, and 20% are exposed to drug B. Sex, drug A, and drug B are independent. All individuals have a baseline risk of disease Y of 5% throughout follow up, men who are exposed to drug A have a 15% increased risk of disease Y, and so do women who are exposed to drug B. If we were using a real-life dataset instead of a simulated one, we would need methods that could help us to identify the different risks associated with the measured exposures and population characteristics in order to eventually prevent disease Y. We will show how the CoOL approach can help direct us towards sets of exposures which might have caused our health outcome of interest in this example and in more complex simulations (see Supplementary material).

The Causes of Outcome Learning approach

The CoOL approach is enabled by three major new developments in the fields of machine learning* and causal inference*. Firstly, advances in computing and machine learning allow for identification of complex structures in large datasets. Secondly, there has been a recent breakthrough in understanding why machine learning models produce the results they do (explainable AI such as Layer-Wise Relevance Propagation [LRP] [11–13]). Lastly, by assuming a causal structure of data, models may be interpreted as structural causal models, which allows for causal interpretation.[14] The CoOL approach follows 3 phases (a-c) to ensure the embeddedness in an inductive-deductive scientific process, of which our contribution is specifically related to the computational phase (Figure 1).

a) Pre-computational phase: Propose a causal model using a Directed Acyclic Graph (DAG) of the exposures and the outcome based on prior domain expertise of selected actionable exposures and contextual factors.
b) Computational phase (We provide the R package 'CoOL', see Supplementary information 1 for installation):
    - On a training dataset:
        i. Fit a *non-negative additive model** based on the features from the assumed causal model
        ii. Decompose the risk contributions.

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

iii. Cluster individuals based on the risk contributions.

- On an internal validation dataset:

iv. Ensure the robustness of the findings in an internal validation dataset.

c) Post-computational phase: Based on the new learnings from the computational phase and existing knowledge, develop hypotheses to be assessed in further (intervention) studies on new data. Our approach focuses on common sub-groups with high risks, and thus the presentation of the results can direct the researchers towards insight with potential large public health impact. The scientific process will continue in an inductive-deductive continuum towards inference to the best explanation.

Pre-computational phase on proposing a causal model

Causal structures are commonly depicted with DAGs,[15] which allow for a causal interpretation of associations given a set of *causal assumptions**: exchangeability, positivity, consistency, no measurement error, and no model misspecification.[6]

The intuition behind the assumed causal model is to link exposures to unknown sufficient causes.[16] Figure 2A ideally shows unknown combinations of exposures that cause the health outcome. Figure 2C shows the theoretical DAG, where X_i denotes exposures, SC_j denotes j unknown sets of sufficient causes for the outcome (inspired by the notation by VanderWeele and Robins [16]), and Y denotes the outcome. The U_{SCi} and U denotes different types of unmeasured (including unmeasurable and unknown) causes; U_{SCi} denotes the unmeasured component causes of SC_j, while U denotes unmeasured causes of Y of which we assume all individuals are exposed to. This theoretical DAG avoids making assumptions of the lack of causal effects between exposures and sufficient causes, and the computational steps will aim at reducing these causal effects towards the minimal sets of component causes. The assumed causal model assists in the selection of exposures to include in the model: there are actionable exposures, i.e. those we can intervene on such as drug intake, and contextual factors, which help describe sub-groups in risk. It also helps to decide whether proximal non-actionable exposures should be left out of the model, as they may mediate effects of actionable exposures. Further, the assumed causal model is used for the interpretation of the results because only direct and joint effects are returned.[5]

The model choice affects our causal interpretations,[17] and models for estimating synergistic effects assume positive monotonicity, i.e. exposures either have no effect or always act in the same direction on the outcome.[18,19] The proposed non-negative model (next section) relaxes the monotonicity assumption by letting us explore all directions of exposures on the outcome simultaneously for which effects act independently or synergistically with others (e.g. if there exists exposures that are especially harmful for men and other exposures that are especially harmful for women).

To identify an elevated risk, we need to define a reference *baseline risk*, R^{b+}, of all unmeasured causes assumed to affect all individuals. Given the causal structure is correct, the average effect for being exposed to combination z of the exposures compared with the baseline risk is given by $P(Y_{X_z} = 1) - P(Y = 1)_{baseline}$, which can be

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

estimated as $P(Y = 1|X = x_z) - R^{b+}$, and the average effect of removing one exposure, by setting X_i to $\overline{x_i}$, can be estimated as $P(Y = 1|X_i = \overline{x_i}, X_{!i} = x_z) - P(Y = 1|X_i = x_i, X_{!i} = x_z)$.

As in all studies aiming at causal inference, appropriate adjustment of exposures causing confounding is of great concern.[6] In the CoOL approach, *confounders** are of interest as they could be causes of the outcome and be additionally associated with the exposures or they could be causes of the exposures in addition to being associated with the outcome.[20] By including the factors responsible for confounding, we block spurious associations between exposures and the outcome, which they would otherwise introduce (*the backdoor path criterion**) [14]. Researchers should also consider situations where, e.g. differences in disease definition or surveillance and changes in registrations of exposures are known to have happened over time.[21] Issues with selection or collider bias and measurement bias should equally well be considered.

For our motivating example, we assume that sex, drug A, and drug B do not share a common cause. Ideally, we want to identify the sufficient causes shown in Figure 2B, and the DAG showing our scientific interest can be drawn as in Figure 2D.

Computational phase

Since the number of potential combinations of exposures is large, there is a risk of type 1 errors. To prevent the model from *overfitting** to noise, data is split into a training dataset and a dataset for internal validation. We suggest training until the model converges based on the error function for the training dataset. The findings from the training dataset can be manually confirmed in the yet unseen internal validation dataset before developing hypotheses.

1. Fitting a non-negative model

Various non-negative models may be used - we suggest a *non-negative* additive* single-hidden layer* neural network** (Figure 2E) designed to mimic our assumed causal model (Figure 2C). This model resembles a linear regression model estimating risk differences but with two main modifications. First, the model includes a series of unobserved mediators that can combine the effects of various exposures. We call these unobserved mediators the *synergy-functions**, $S^+()$, represented in the hidden layer between the exposures and the outcome. Second, we restrict all connection parameters to have non-negative values,[22] so that exposures can only increase the occurrence of the outcome and thereby meet a relaxed version of the monotonicity assumption.[18] In Figure 2E, X_i denotes i exposures (each category of the variable is *binary (/one-hot) encoded** into one new variable each with 0 if not present and 1 if present) and Y denotes the disease outcome (coded 0 and 1); $\beta^+_{i,j}$ denotes connection parameters from the exposures to the synergy-functions (parameters can only take non-negative values), $S^+()$, which return the non-negative sum of its input value or zero; α^-_j is an intercept (can only take non-positive values) that acts as an activation threshold which only allows combinations of exposures with large $\beta^+_{i,j}$-weighted sum to pass $S^+()$; R^{b+} represents the baseline risk (can only take non-negative values); and the parameter of connections between the synergy functions and the outcome has a fixed value of 1. The model

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

estimates the risk on an additive scale so that synergism is defined as combined effects that are larger than the sum of individual effects.[2] Clearly, probabilities larger than 1 indicate model misspecification.

This model can be denoted as below, where $S^+(x) = max(0, x)$, and likewise $^+$ denotes restrictions to non-negative values, and $^-$ denotes restrictions to non-positive values. This model satisfies the assumption that the added risk is independent of the baseline risk or phrased as an "independent of background" model by Beyea and Greenland.[17]

$$P(Y = 1|X) = \sum_j \left(S^+ \left(\sum_i (X_i \cdot \beta_{i,j}^+) + \alpha_j^- \right) \right) + R^{b+}$$

Fitting the model is done using *stochastic gradient decent** on the training dataset: In a step-wise procedure run on one individual at a time, the model estimates the individual's risk of the disease outcome, $P(Y|X)$, calculates the squared prediction error $(Y - P(Y|X))^2$ and adjusts the model parameters to minimize this error.[23] By iterating through all individuals for multiple *epochs**, we obtain model parameters, which minimizes the sum of prediction errors across the entire population. The *initial values**, *derivatives**, *learning rates**, and *regularizations** are described in Supplementary information 2.

While the prediction performance measured by the area under the receiver operating characteristic curve (AUC) provides a useful metric for evaluating model discriminatory performance across the entire population, it is important to consider that in case the outcome is caused by multiple distinct sets of causes, a model with low AUC can still capture sets of causes for a particular sub-group.[24]

Figure 2F shows the model for our motivating example. We binary encode new variables for each possible category of each exposure, such that sex (coded 0 if man, 1 if women) becomes two factors; man (coded 1 if man, 0 if not man) and woman (coded 1 if woman, 0 if not woman) and so forth for drug A and drug B. This data is used to fit the proposed non-negative model with 10 synergy-functions. Figure 4A-C shows how the error decreases by each epoch, it visualizes the neural network connections, and shows the accuracy using the prediction performance measure, AUC.

2. Decomposing risk contributions

Machine learning models are commonly referred to as black boxes due to the limited interpretability of their parameters and how they interact with the input-variables.[25] Instead of attempting to interpret the model directly, we use LRP [11–13] to decompose the risk of the outcome to *risk contributions** for each individual (in particular, we use the $LRP_{alpha,beta}$-rule, with alpha=1 and beta=0). LRP was introduced by Bach et al. in 2015[11] as a decomposition technique for pre-trained neural networks, and was later justified via Deep Taylor Decomposition.[26] As opposed to other explanation techniques for neural networks, LRP is aimed at conserving the information for predicting the outcome when assigning relevance to the inputs that were driving the prediction. In the CoOL approach, the predicted risk of the outcome, $P(Y = 1|X)$ is decomposed into a baseline risk, R^{b+}, and the risk contributions by each exposure, R_i^X (where $P(Y = 1|X)$ can take values between 0 and 1):

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

$$R^{b+} + \sum_i R_i^X = P(Y = 1|X)$$

These risk contributions may be interpreted as the exposures' positive contribution to the risk given the model and the individual's set of exposures. No risk contributions are decomposed to the intercepts, α_j^-. The below procedure is conducted for all individuals in a one-by-one fashion. The baseline risk, R^{b+}, is represented by its own parameter as illustrated in Figure 2E, and is therefore estimated as part of fitting the non-negative neural network. The decomposition of the risk contributions for exposures, R_i^X, takes 3 steps:

Step 1 - Subtract the baseline risk, R^{b+}:

$$R_{total}^X = P(Y = 1|X) - R^{b+}$$

Step 2 - Decompose risk contributions to the synergy-functions, where S_j is the value returned by each of the j synergy-functions given the exposure distribution X_i, parameters, $\beta_{i,j}^+$, and intercepts, α_j^-:

$$R_j^X = \frac{S_j}{\sum_{j'} S_{j'}} R_{total}^X$$

Step 3 - Decompose risk contributions from the synergy-functions to the exposures:

$$R_i^X = \sum_j \left(\frac{X_i \cdot \beta_{i,j}^+}{\sum_{i'} (X_{i'} \cdot \beta_{i',j}^+)} R_j^X \right)$$

As a result of the risk decomposition, each individual is assigned a set of risk contributions, R_i^X, one for each exposure plus a baseline risk, R^{b+}. The decomposition of risk contributions can be illustrated in Figure 3E-F using the motivating example and explanation in the figure legend.

3. Clustering of risk contributions

We suggest to sub-group the individuals based on risk contributions using *Manhattan distances** and *the Ward method**.[27,28] One helpful technique that could inform deciding on number of subgroups is a dendrogram[29] of the distance matrix with node sizes representing the prevalence of similar risk contributions to be used to decide the number of sub-groups (Figure 3G). Additionally, we can plot the prevalence and mean risk of each sub-groups (inspired by excess probability plots[30]) to help identify the sub-groups with a high impact by identifying the area within a sub-group above the baseline risk (Figure 3I). Further, we can make a table of mean risk contributions and standard deviations (SD) by sub-groups to illuminate which exposures elevate the risk in each sub-group (Figure 3J). An indication of synergism is when the combined risk contribution of a set of exposures is higher than the sum of stand-alone risk contributions of each of the exposures, which can also be added to the table (Supplementary information 3, but it should be interpreted with caution as deviation may occur in noisy datasets). Formal investigation of synergism should be done in the yet unseen internal validation dataset before developing hypotheses for phase 3 of the Causes of Outcome Learning approach (see formulas for formal interaction analyses by VanderWeele[31]).

Pre-print (not peer-reviewed or accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

If exposures identified as high risk contributors are associated with the outcome due to uncontrolled confounding, the risk contributors only help identify the high-risk group and removing the exposure will not necessarily affect the risk. If two component causes act solely in synergy on the outcome, then removing just one of them is sufficient by itself, and thus the estimated risk contributions underestimate the causal effects in a counterfactual framework. The area within a sub-group above the baseline risk (Figure 3I) indicate the excess fraction of all cases due to the combination of exposures in the sub-group and thus indicate groups where a large public health impact may be made, but the interpretation should depend on how well the causal mechanisms are understood. The term also relates to the concept of grouped partial attributable risks[32] or termed formally as the attributable proportion in the population[19] and can be defined for a subgroup Z as:

$$\frac{P(Y=1) - P(Y_{X_z=\overline{X_z}}=1)}{P(Y=1)}$$

where $X_z = \overline{X_z}$ denotes eliminating risk contributors in subgroup Z. Given the combined risk contributions causally affect the outcome and meet the assumption of positive monotonicity, the excess fraction can be calculated as (Supplementary information 4):

$$\frac{P(X=x_z)\cdot\left(P\left(Y_{X_z}=1\right) - R^{b+}\right)}{P(Y=1)}$$

Analyzing our motivating example, we can apply the fitted non-negative model, decompose the risk contributions using LRP and show a dendrogram of how similar the population are to each other in Figure 4D, which indicated 3 groups. Figure 4E shows the risk and prevalence of the 3 sub-groups, where one sub-group which has a risk of 5%, a second sub-group that has a risk of approximately 20% with a prevalence of 10%, and a third sub-group that has a risk of approximately 20% with a prevalence of 10%. Figure 4F shows us that we correctly identified that men (sex_0) who are exposed to drug A (drug_a_1) have a 5% baseline risk, which reaches a near 20% risk through the contributions from being a man and drug A. Similar are the findings for women (sex_1) and drug B (drug_b_1). Though not observed in this analysis, we may expect that the predicted risks are slightly underestimated since we apply regularization to reduce noise signals in data.

Post-computational phase on hypothesis development and validation

The results of the computational step may provide learnings about different sets of exposures, which may have led to the outcome in different sub-populations. This evidence should be interpreted in the light of the assumed causal model that was specified in phase 1, and thus formulated into new hypotheses about multifactorial etiology, which may be denoted in a DAG as done by VanderWeele.[16] The empirical evidence from the computational phase highlights the outcome prevalence and risk distribution across population sub-groups and directs attention towards groups with a potentially large public health impact.

The domain experts will need to assess whether the causal assumptions are met across the identified sub-groups. Hypothetically, it may be that unmeasured confounding influenced our results based on our prior causal assumptions, which suggests that further work needs to be conducted to validate the findings and to understand

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

how risk may be mitigated for the sub-groups. The major gain of using the CoOL approach compared to many other machine learning approaches is that the sub-groups can be defined by specific combinations of exposures that are easily communicated with words rather than by a black box algorithm. Hence, our learnings may be formulated as a hypothetical intervention and explored using established methodological frameworks for causal inference modelling.[6] Here, we can use the frameworks developed for synergistic effects of causes in a causal inference framework to draw our causal assumptions.[16]

Phase 3 for triangulating the hypotheses is conducted in external populations (either temporal validation or more desirably, external validation). If replicable, the researchers should provide sufficient evidence that the replicated finding is causal (and not due to similar bias structures) for example using various triangulation approaches with orthogonal bias structures (i.e. designs with biases in different directions) including studies outside the epidemiological field.[33] Eventually, if possible, the hypotheses needs to be tested using a randomized set-up.

In our example, we now have some learnings to inform two hypotheses: Men taking drug A seem to be at a higher-than-normal risk, women taking drug B seem to be at a higher-than-normal risk. We may supplement with observational data from other settings before we eventually may intervene (stop exposure to drug A for men, and drug B for women) possible in a randomized way if justified by *equipoise**.

Discussion

We have introduced the CoOL approach, which investigates common combinations of exposures, which may have led to a specific health outcome. We have used a simple simulated example in the presentation, however, the approach applies to more complex scenarios (see additional simulations and a real life data analysis in the Supplementary material). New learnings can be formulated as hypotheses in words rather than a black box algorithm, and these hypotheses can subsequently be challenged and tested using for example the framework for hypothetical interventions and by triangulation.

So far, the sufficient cause model and the way of thinking about causes of an outcome have, *de facto*, mostly been a theoretical framework and not a practical approach for applied data analysis in epidemiology.[6] Though, there exists justifications of individual (n=1) explanation rather than group-based explanations by going from studying effects of causes to causes of an outcome,[34] our intention with the CoOL approach is to identify commonly shared sets of events, which are associated with higher risks of the outcome in specific sub-populations for public health interventions. Fully explaining an outcome seems far-fetched in epidemiology,[35] since these sets of events will interplay with multiple unknown or unmeasurable causes, but an approach like ours takes the first steps towards suggesting etiology[36] or – at least – to identify vulnerable subgroups.

The proposed approach may be of relevance to a number of theoretical frameworks: It links to the classical sufficient causes model[2], it may help disentangle structures in the *syndemics* (synergistic epidemics) literature,[37] and add a tool for holistic approaches to "precision" public health.[38] We stress that the CoOL

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

approach is an inductive-deductive approach and that researchers in each of the phases need to carefully consider the most appropriate set-up that eventually may lead to fair public health actions.

Limitations and extensions

Inference

Co-occurring associations between the exposures and the outcome can be due to various causal structures such as interactions,[31] clustered causes (exposures sharing a common cause), mediation,[31] uncontrolled confounding,[6] and conditioning on a common effect (collider-stratification or selection bias)[39] (Figure 5). *Interactions* (Figure 5A) entail a combined structural effect that is beyond the sum of the individual effects of the putative causes, and thus some inference about the underlying structures may be suggested by the CoOL approach and confirmed by formal interaction analysis. When studying non-randomized epidemiological data, complex combinations of all of these structures can be expected. Researchers will need to assess various hypotheses through triangulation to support their explanatory contribution.

Rose described chains of causes by separating causes into distal and proximal causes.[3] Proximal causes, e.g. infectious agents, dietary deficiencies, smoking, toxic exposures and allergens, are close to the outcome in the causal chain, and distal causes, e.g. social and economic positions, as the causes of causes and thus are distal to the outcome in the causal chain. Such frameworks have been further expanded in the exposome literature.[40] The CoOL approach focuses on the most proximal causes of the causal chain, and thus included exposures should be carefully selected according to appropriate actionable exposures and contextual factors. An individualized focus on proximal causes may misdirect our attention away from structural public health interventions and could in the worst case scenario stigmatize parts of the population without offering preventive interventions.[41] Future work is needed to explore the degree of which bias is introduced due to collider bias by using a neural network.[16]

Model

The version of CoOL we have presented deals with binary exposures and outcomes, similar to the sufficient cause model.[7] However, the approach can be extended to continuous outcomes, where the value 0 has a meaningful interpretation (as e.g. loss of disease-free years, and in contrast to e.g. body mass index). Further, it may be that the CoOL approach reveals complementary information when studying positive outcomes (high quality of life) in comparison to negative outcomes (diseases and death).[42] Multiple other extensions of the CoOL approach may be possible, e.g. ways to incorporate time, such as time-varying variables, complex confounding scenarios, and censoring, would be of high relevance for epidemiological, and we encourage others to explore these. It could be of interest to explore a variation of outcome-wide approaches,[43] since co-morbidity may be a sign of shared underlying sufficient component causes (e.g. atopic diseases as asthma, dermatitis and nasopharyngitis), and thus analyzing several health outcomes may let us pick up suggestions of common sufficient causes otherwise missed (*multi-task learning**). The best presentation of the results will depend on the aim and extensions of the CoOL approach.

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Robustness checks

If data are sparse for investigating the causal structure, the results may not be reliable both with the potential of type 1 (false positive findings) and type 2 errors (false negative findings). Robustness checks should be conducted to challenge the stability of the approach.[44] It can give insight to try the following to ensure that the baseline risk estimation and the identified groups are robust: change the number of synergy-functions (Supplementary simulation 7) re-run the analysis with sub-samples of the study population (Supplementary simulation 8), change the regularization values (Supplementary simulation 9). A baseline risk above its initial value, $\frac{\sum Y}{n}$, is a sign of model misspecification.

Theoretical comparison with other approaches

Though not commonly applied in epidemiology, frameworks for identifying component causes exist by selection on either cases[9]or exposure[10]. In the social sciences, Configurational Comparative Methods deal with sufficient causes (referencing earlier work[45]), of which the most famous is the Qualitative Comparative Analysis,[46] which has also been applied in the public health domain.[47] Qualitative comparative analysis works by analyzing all combinations of exposures and uses a top-down search of exposure combinations which fulfill some chosen criteria, such as a risk threshold.[8] Using pre-defined risk thresholds may both have advantages such as transparent protocols and disadvantages such as being sensitive to the chosen threshold level with the risk of not identifying relevant risk groups with moderate increased risks but with large public health impact.

The CoOL approach has similarities to decomposition approaches of mediated and interactive effects in epidemiology,[48,49] however, work is needed to assess the similarities to the LRP properties in the CoOL approach. A recent approach, Algorithm for Learning Pathway Structures (ALPS), which uses a Monte Carlo scheme to update a pathway structure has shown promise for identifying complex interactions in large epidemiological datasets.[50] However, since ALPS focuses on parameter interpretation, its results differs from those of the CoOL approach, which identifies sizeable sub-groups who share exposures, which may have led to their increased risk of the outcome.

Approaches such as the exposome[40,51,52] and exposure- or environment-wide association studies (EWAS)[53,54] assess multiple exposures simultaneously, but few applied studies include interactions.[53–56] The few that do consider interactions tend to investigate interaction of pre-selected factors only.[57] Such studies have been discussed in relation to their potential, especially in light of successes of genome-wide association studies,[58] and limitations such as a challenging causal interpretation.[5]

LRP has previously been successfully demonstrated in image, text and biological data classification,[59–61] as well as for health records to explain clinical decisions on therapy assignment.[62] In this latter case, neither a baseline risk was estimated nor was there interest in identifying sub-groups. The computational phase of the CoOL approach has similarities to existing work on explaining and correcting computer vision,[63,64] but takes its depart from a causal question. The CoOL approach may be viewed as a supervised clustering approach based

Preprint (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

on an additive feature attribution method guided by a causally inspired model. It should be investigated to which degree other additive feature attribution methods approximate similar results,[65] since they are usually not related to any causal frameworks nor any underlying epidemiological theories. These methods seem to depart from a specified reference group in contrast to the CoOL approach, which estimates a baseline risk. Different model dependent methods for decomposing risks can naturally yield different estimates.[66]

Alternative methods to LRP for decomposing neural network predictions were proposed recently, such as DeepLIFT[67] or Integrated Gradients.[68] However, only LRP and its Deep Taylor Decomposition theoretical framework[26] fit our assumption of a non-negative neural network with negative biases, and allow for a seamless interpretation of relevance as risk contributions in our causal inference setup. Using non-negative models for sets of explanations within certain aims was proposed decades ago[69] but not in relation to causal questions. We did not want to consider sensitivity-, perturbation-, or surrogate-based explanation techniques, since our question of interest relates to the causes of an outcome posed as "*Given a particular health outcome, what are the most common sets of exposures, which might have been its causes?*" rather than effects of causes posed as "*What would have occurred if a particular factor were intervened upon and thus set to a different level than it in fact was?*" These distinctions have previously been discussed both in the causal inference literature[7] and in the literature on LRP.[12] Furthermore, perturbation-based methods produce localized explanations which may not generalize to global causal pathways.[59,70,71]

Conclusion

We have introduced the Causes of Outcome Learning approach with the aim of disentangling common combinations of pre-outcome exposures that could have caused a specific health outcome. The approach is based on prior knowledge of the causal structure, the flexibility of a non-negative neural network, the LRP explanation technique for decomposing risk contributions and clustering, and, finally, hypothesis development and testing. These are steps towards building better transparency and causal reasoning into hypothesized causal findings from machine learning methods in the health sciences.[72,73] This CoOL approach should encourage and enable epidemiologists to examine common combinations of exposures as causes of the outcome of interest. This could eventually inform the development of more effective, targeted and impactful public health interventions.

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

References

1 Ngamwong Y, Tangamornsuksan W, Lohitnavy O, *et al.* Additive synergism between asbestos and smoking in lung cancer risk: A systematic review and meta-analysis. PLoS One. 2015. doi:10.1371/journal.pone.0135798

2 Rothman KJ. Causes. *Am J Epidemiol* 1976;**104**:587–92. doi:10.1093/oxfordjournals.aje.a112335

3 Rose G, Khaw KT, Marmot M. *Rose's Strategy of Preventive Medicine*. 2009. doi:10.1093/acprof:oso/9780192630971.001.0001

4 Brankovic M, Kardys I, Steyerberg EW, *et al.* Understanding of interaction (subgroup) analysis in clinical trials. Eur. J. Clin. Invest. 2019. doi:10.1111/eci.13145

5 Vanderweele TJ. Outcome-wide epidemiology. *Epidemiology* Published Online First: 2017. doi:10.1097/EDE.0000000000000641

6 Hernan M, Robins J. *Causal Inference*. Boca Raton: Chapman & Hall/CRC, forthcoming. 2018.

7 VanderWeele TJ, Hernán MA. From counterfactuals to sufficient component causes and vice versa. Eur. J. Epidemiol. 2006. doi:10.1007/s10654-006-9075-0

8 Ragin CC. Using qualitative comparative analysis to study causal complexity. *Health Serv Res* 1999.

9 Reiber GE, Vileikyte L, Boyko EJ, *et al.* Causal pathways for incident lower-extremity ulcers in patients with diabetes from two settings. *Diabetes Care* Published Online First: 1999. doi:10.1001/archfami.3.3.273

10 Alrawahi AH. New approaches to disease causation research based on the sufficient-component cause model. *J Public health Res* Published Online First: 2020. doi:10.4081/jphr.2020.1726

11 Bach S, Binder A, Montavon G, *et al.* On pixel-wise explanations for non-linear classifier decisions by layer-wise relevance propagation. *PLoS One* Published Online First: 2015. doi:10.1371/journal.pone.0130140

12 Montavon G, Samek W, Müller KR. Methods for interpreting and understanding deep neural networks. Digit. Signal Process. A Rev. J. 2018. doi:10.1016/j.dsp.2017.10.011

13 Montavon G, Binder A, Lapuschkin S, *et al.* Layer-Wise Relevance Propagation - an Overview. In: *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. 2019. 193–206. doi:10.1007/978-3-030-28954-6_10

14 Pearl J. Causal Inference in Statistics: A Primer - Judea Pearl, Madelyn Glymour, Nicholas P. Jewell - Google Bøger. https://books.google.dk/books/about/Causal_Inference_in_Statistics.html?id=IqCECwAAQBAJ&redir_esc=y (accessed 12 Aug 2020).

15 Tennant PW, Harrison WJ, Murray EJ, *et al.* Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. *medRxiv* 2019.

Preprint (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

16 Vanderweele TJ, Robins JM. Directed acyclic graphs, sufficient causes, and the properties of conditioning on a common effect. *Am J Epidemiol* 2007;**166**:1096–104. doi:10.1093/aje/kwm179

17 Beyea J, Greenland S. The importance of specifying the underlying biologic model in estimating the probability of causation. *Health Phys* Published Online First: 1999. doi:10.1097/00004032-199903000-00008

18 VanderWeele TJ, Robins JM. The identification of synergism in the sufficient-component-cause framework. *Epidemiology* Published Online First: 2007. doi:10.1097/01.ede.0000260218.66432.88

19 Suzuki E, Yamamoto E, Tsuda T. On the relations between excess fraction, attributable fraction, and etiologic fraction. *Am J Epidemiol* Published Online First: 2012. doi:10.1093/aje/kwr333

20 Arah OA. Bias Analysis for Uncontrolled Confounding in the Health Sciences. *Annu Rev Public Health* Published Online First: 2017. doi:10.1146/annurev-publhealth-032315-021644

21 Rieckmann A, Nguyen T-L, Dworzynski P, *et al.* Machine Learning models aimed at identifying risk factors for reducing morbidity and mortality may need to consider confounding such as calendar time. *Submitted*

22 Kallus N. Classifying Treatment Responders Under Causal Effect Monotonicity. 2019.

23 LeCun YA, Bottou L, Orr GB, *et al.* Efficient BackProp BT - Neural Networks: Tricks of the Trade. In: *Neural Networks: Tricks of the Trade*. 2012. doi:10.1007/978-3-642-35289-8_3

24 Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int J Epidemiol* Published Online First: 2020. doi:10.1093/ije/dyz274

25 Pearl J. The seven tools of causal inference, with reflections on machine learning. *Commun ACM* 2019;**62**:54–60. doi:10.1145/3241036

26 Montavon G, Lapuschkin S, Binder A, *et al.* Explaining nonlinear classification decisions with deep Taylor decomposition. *Pattern Recognit* Published Online First: 2017. doi:10.1016/j.patcog.2016.11.008

27 Strauss T, Von Maltitz MJ. Generalising ward's method for use with manhattan distances. *PLoS One* Published Online First: 2017. doi:10.1371/journal.pone.0168288

28 Chavent M, Kuentz-Simonet V, Labenne A, *et al.* ClustGeo: an R package for hierarchical clustering with spatial constraints. *Comput Stat* Published Online First: 2018. doi:10.1007/s00180-018-0791-1

29 Yu G, Smith DK, Zhu H, *et al.* ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* Published Online First: 2017. doi:10.1111/2041-210X.12628

30 Eide GE, Heuch I. Attributable fractions: fundamental concepts and their visualization. *Stat Methods Med Res* 2001;**10**:159–93. doi:10.1177/096228020101000302

31 VanderWeele TJ. *Explanation in Causal Inference: Methods for Mediation and Interaction*. Oxford University

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Press 2015.

32 Land M, Vogel C, Gefeller O. Partitioning methods for multifactorial risk attribution. Stat. Methods Med. Res. 2001. doi:10.1191/096228001680195166

33 Lawlor DA, Tilling K, Smith GD. Triangulation in aetiological epidemiology. *Int J Epidemiol* Published Online First: 2016. doi:10.1093/ije/dyw314

34 Pearl J. Causes of Effects and Effects of Causes. *Sociol Methods Res* Published Online First: 2015. doi:10.1177/0049124114562614

35 Smith GD. Epidemiology, epigenetics and the 'Gloomy Prospect': Embracing randomness in population health research and practice. *Int J Epidemiol* Published Online First: 2011. doi:10.1093/ije/dyr117

36 Olsen J. What characterises a useful concept of causation in epidemiology? J. Epidemiol. Community Health. 2003. doi:10.1136/jech.57.2.86

37 Tsai AC. Syndemics: A theory in search of data or data in search of a theory? Soc. Sci. Med. 2018;**206**:117–22. doi:10.1016/j.socscimed.2018.03.040

38 Olstad DL, McIntyre L. Reconceptualising precision public health. *BMJ Open* 2019;**9**:e030279. doi:10.1136/bmjopen-2019-030279

39 Arah OA. Analyzing Selection Bias for Credible Causal Inference. *Epidemiology* Published Online First: 2019. doi:10.1097/ede.0000000000001033

40 Wild CP. The exposome: From concept to utility. Int. J. Epidemiol. 2012. doi:10.1093/ije/dyr236

41 Kee F, Taylor-Robinson D. Scientific challenges for precision public health. J. Epidemiol. Community Health. 2020. doi:10.1136/jech-2019-213311

42 Vanderweele TJ, Chen Y, Long K, *et al.* Positive Epidemiology? Epidemiology. 2020. doi:10.1097/EDE.0000000000001147

43 VanderWeele TJ, Mathur MB, Chen Y. Outcome-Wide Longitudinal Designs for Causal Inference: A New Template for Empirical Studies. *Stat Sci* Published Online First: 2020. doi:10.1214/19-sts728

44 Lange T, Roth V, Braun ML, *et al.* Stability-based validation of clustering solutions. *Neural Comput* Published Online First: 2004. doi:10.1162/089976604773717621

45 Mackie JL. Causes and Conditions. *Am Philos Q* 1965;**2**:245–64.

46 Baumgartner M. Configurational causal modeling and logic regression.

47 Warren J, Wistow J, Bambra C. Applying qualitative comparative analysis (QCA) in public health: A case study of a health improvement service for long-term incapacity benefit recipients. *J Public Heal (United Kingdom)* Published Online First: 2014. doi:10.1093/pubmed/fdt047

Pre-print (not peer reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

48 Vanderweele TJ. A three-way decomposition of a total effect into direct, indirect, and interactive effects. *Epidemiology* Published Online First: 2013. doi:10.1097/EDE.0b013e318281a64e

49 Huang YT, Tai AS, Chou MY, *et al.* Six-way decomposition of causal effects: Unifying mediation and mechanistic interaction. *Stat Med* Published Online First: 2020. doi:10.1002/sim.8708

50 Baurley J, Kjærsgaard A, Zwick M, *et al.* Bayesian Pathway Analysis for Complex Interactions. *Am J Epidemiol* Published Online First: 2020. doi:10.1093/aje/kwaa130

51 Rappaport SM, Smith MT. Environment and disease risks. Science (80-.). 2010. doi:10.1126/science.1192603

52 Rappaport SM. Implications of the exposome for exposure science. J. Expo. Sci. Environ. Epidemiol. 2011. doi:10.1038/jes.2010.50

53 Patel CJ, Bhattacharya J, Butte AJ. An environment-wide association study (EWAS) on type 2 diabetes mellitus. *PLoS One* Published Online First: 2010. doi:10.1371/journal.pone.0010746

54 Patel CJ, Bhattacharya J, Ioannidis JPA, *et al.* Systematic identification of correlates of HIV infection: An X-wide association study. *AIDS* Published Online First: 2018. doi:10.1097/QAD.0000000000001767

55 Tzoulaki I, Patel CJ, Okamura T, *et al.* A nutrient-wide association study on blood pressure. *Circulation* Published Online First: 2012. doi:10.1161/CIRCULATIONAHA.112.114058

56 Patel CJ, Cullen MR, Ioannidis JP, *et al.* Systematic evaluation of environmental factors: Persistent pollutants and nutrients correlated with serum lipid levels. *Int J Epidemiol* Published Online First: 2012. doi:10.1093/ije/dys003

57 Patel CJ, Ioannidis JPA, Cullen MR, *et al.* Systematic assessment of the correlations of household income with infectious, biochemical, physiological, and environmental factors in the United States, 1999-2006. *Am J Epidemiol* Published Online First: 2015. doi:10.1093/aje/kwu277

58 Ioannidis JPA. Exposure-wide epidemiology: Revisiting Bradford Hill. *Stat Med* Published Online First: 2016. doi:10.1002/sim.6825

59 Samek W, Binder A, Montavon G, *et al.* Evaluating the visualization of what a Deep Neural Network has learned. *IEEE Trans Neural Networks Learn Syst* 2015;**28**:2660–73. doi:10.1109/TNNLS.2016.2599820

60 Arras L, Horn F, Montavon G, *et al.* 'What is relevant in a text document?': An interpretable machine learning approach. *PLoS One* 2017;**12**. doi:10.1371/journal.pone.0181142

61 Sturm I, Lapuschkin S, Samek W, *et al.* Interpretable Deep Neural Networks for Single-Trial EEG Classification. *J Neurosci Methods* Published Online First: 2016. doi:10.1016/j.jneumeth.2016.10.008

62 Yang Y, Tresp V, Wunderle M, *et al.* Explaining therapy predictions with layer-wise relevance propagation in neural networks. In: *Proceedings - 2018 IEEE International Conference on Healthcare Informatics, ICHI 2018*. 2018. doi:10.1109/ICHI.2018.00025

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

63 Lapuschkin S, Wäldchen S, Binder A, *et al.* Unmasking Clever Hans predictors and assessing what machines really learn. *Nat Commun* Published Online First: 2019. doi:10.1038/s41467-019-08987-4

64 Anders CJ, Marinč T, Neumann D, *et al.* Analyzing ImageNet with Spectral Relevance Analysis: Towards ImageNet un-Hans'ed. *arXiv* Published Online First: 22 December 2019.http://arxiv.org/abs/1912.11425

65 Lundberg SM, Lee SI. A unified approach to interpreting model predictions. In: *Advances in Neural Information Processing Systems*. 2017.

66 Rabe C, Lehnert-Batar A, Gefeller O. Generalized approaches to partitioning the attributable risk of interacting risk factors can remedy existing pitfalls. *J Clin Epidemiol* Published Online First: 2007. doi:10.1016/j.jclinepi.2006.06.024

67 Shrikumar A, Greenside P, Kundaje A. Learning important features through propagating activation differences. In: *34th International Conference on Machine Learning, ICML 2017*. 2017.

68 Sundararajan M, Taly A, Yan Q. Axiomatic attribution for deep networks. In: *34th International Conference on Machine Learning, ICML 2017*. 2017. http://proceedings.mlr.press/v70/sundararajan17a.html

69 Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature* Published Online First: 1999. doi:10.1038/44565

70 Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. European Conference on Computer Vision (ECCV 2014) pp. 818-833 2014. doi:10.1007/978-3-319-10590-1_53

71 Fong RC, Vedaldi A. Interpretable Explanations of Black Boxes by Meaningful Perturbation. In: *Proceedings of the IEEE International Conference on Computer Vision*. 2017. doi:10.1109/ICCV.2017.371

72 Beam AL, Manrai AK, Ghassemi M. Challenges to the Reproducibility of Machine Learning Models in Health Care. JAMA - J. Am. Med. Assoc. 2020. doi:10.1001/jama.2019.20866

73 Holzinger A, Langs G, Denk H, *et al.* Causability and explainability of artificial intelligence in medicine. Wiley Interdiscip. Rev. Data Min. Knowl. Discov. 2019. doi:10.1002/widm.1312

74 Vollset SE. Confidence intervals for a binomial proportion. *Stat Med* Published Online First: 1993. doi:10.1002/sim.4780120902

Pre-print (not peer-reviewed nor accepted by a journal)

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

Figures

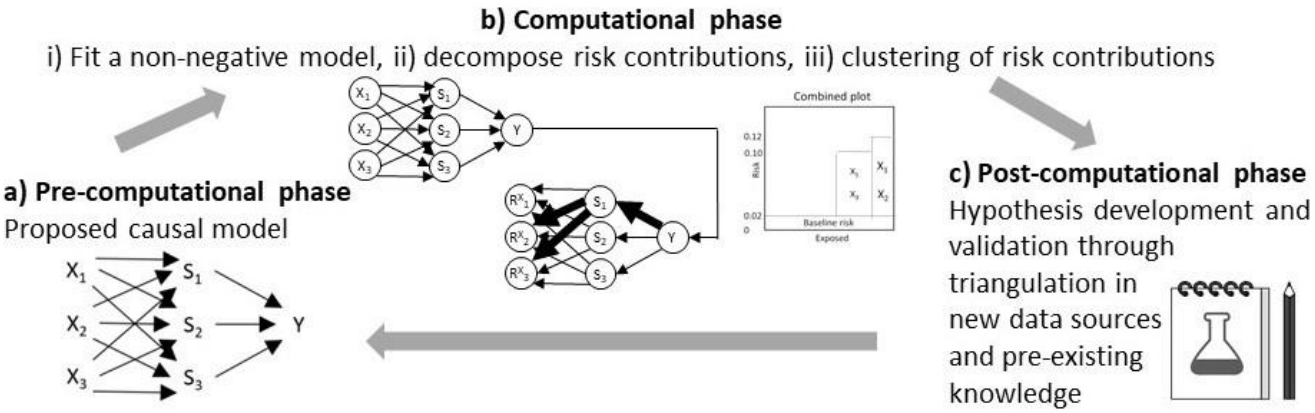


Figure 1. The phases of the CoOL approach towards inference to the best explanation

A) PRE-COMPUTATIONAL PHASE: SCOPING THE RESEARCH QUESTION AND CAUSAL STRUCTURE ASSUMPTIONS. COMPUTATIONAL PHASE: I) A NON-NEGATIVE MODEL AS CLOSE TO THE ASSUMED CAUSAL MODEL IS FITTED, II) RISK CONTRIBUTIONS ARE DECOMPOSED AND III) INDIVIDUALS ARE CLUSTERED INTO SUB-GROUPS. IV) MANUAL VALIDATION OF THE RESULTS IS SUGGESTED IN AN INTERNAL VALIDATION DATASET TO ASSESS THE STABILITY OF THE RESULTS. C) POST-COMPUTATIONAL PHASE: THE RESULTS ARE HELD AGAINST EXISTING EVIDENCE IN ORDER TO DEVELOP NEW HYPOTHESES THAT CAN BE TESTED IN NEW STUDIES. NEW UNDERSTANDINGS WILL UPDATE OUR INITIAL ASSUMED CAUSAL MODEL.

Pre-print (not peer-reviewed nor accepted by a

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

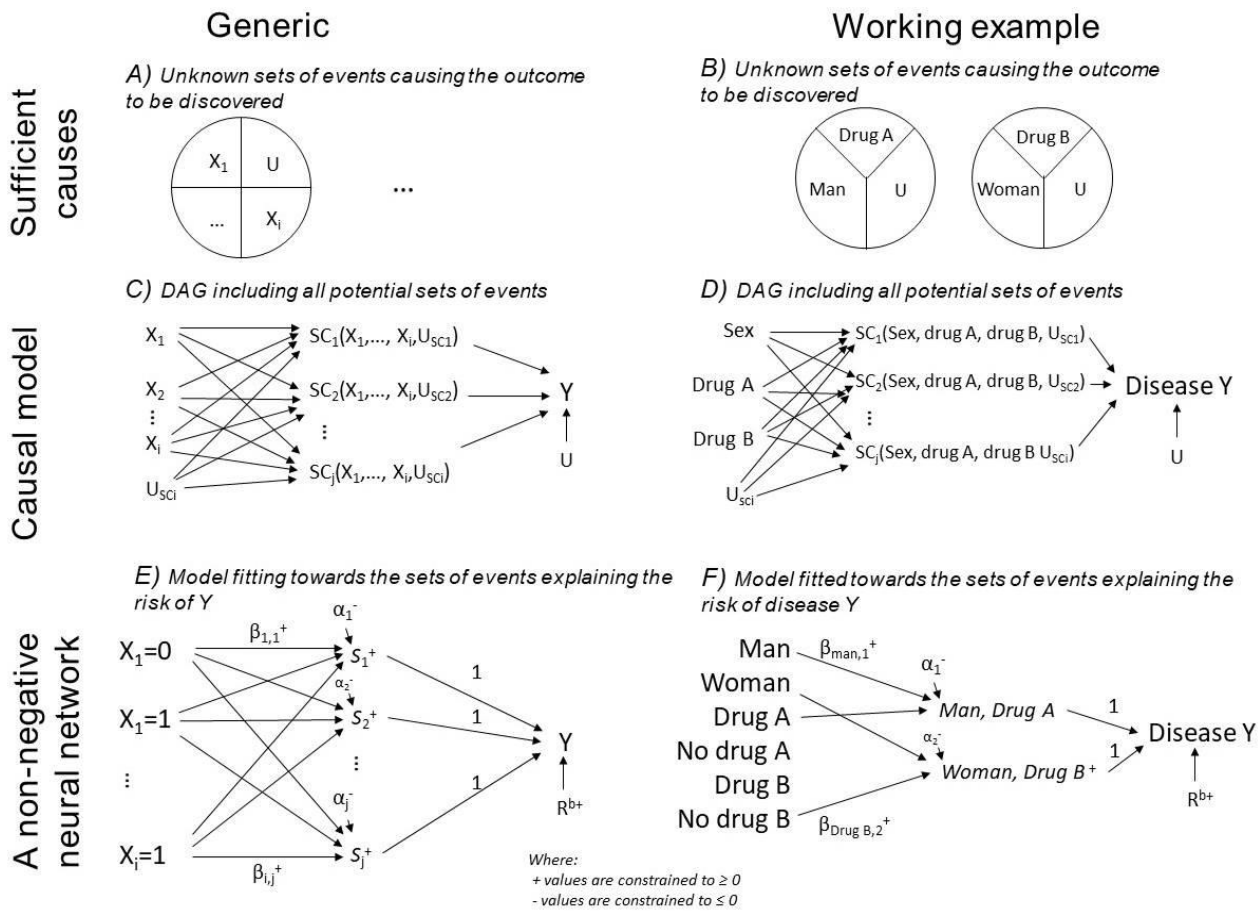


Figure 2. Sufficient causes, causal model and non-negative neural network

THE PICTOGRAM SHOWS THE RELATION BETWEEN EPIDEMIOLOGICAL THEORY, STRUCTURAL MODELS AND A NON-NEGATIVE NEURAL NETWORK. THE LEFT COLUMN IS A GENERIC PRESENTATION, AND THE RIGHT COLUMN SHOWS THE SIMULATED EXAMPLE. A-B) AN ILLUSTRATION OF SUFFICIENT CAUSES. THE EXAMPLE TO THE RIGHT SHOWS THAT A CERTAIN DISEASE OCCURS IF MEN ARE EXPOSED TO DRUG A AND SOME UNKNOWN FACTORS AND IF WOMEN ARE EXPOSED TO DRUG B AND SOME UNKNOWN FACTORS. C-D) AN ASSUMED CAUSAL MODEL ILLUSTRATED USING A DIRECTED ACYCLIC GRAPH, WHERE X_i DENOTES THE EXPOSURES, U_{SCI} DENOTES THE UNMEASURED CAUSES OF THE SUFFICIENT CAUSES, U DENOTES THE UNMEASURED CAUSES OF Y ASSUMED TO AFFECT ALL INDIVIDUALS, SC_j DENOTES *HIDDEN* SUFFICIENT CAUSES, AND Y DENOTES THE OUTCOME. E-F) A NON-NEGATIVE NEURAL NETWORK RESEMBLING THE ASSUMED CAUSAL MODEL. X_i DENOTES EXPOSURES, $\beta^+_{i,j}$ DENOTES NON-NEGATIVE PARAMETERS, S^+_j DENOTES HIDDEN SYNERGY-FUNCTIONS, α^-_j DENOTES NON-POSITIVE INTERCEPTS, ACTING AS ACTIVATION THRESHOLDS FOR SYNERGY-FUNCTIONS, AND R^{b^+} DENOTES THE BASELINE RISK.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

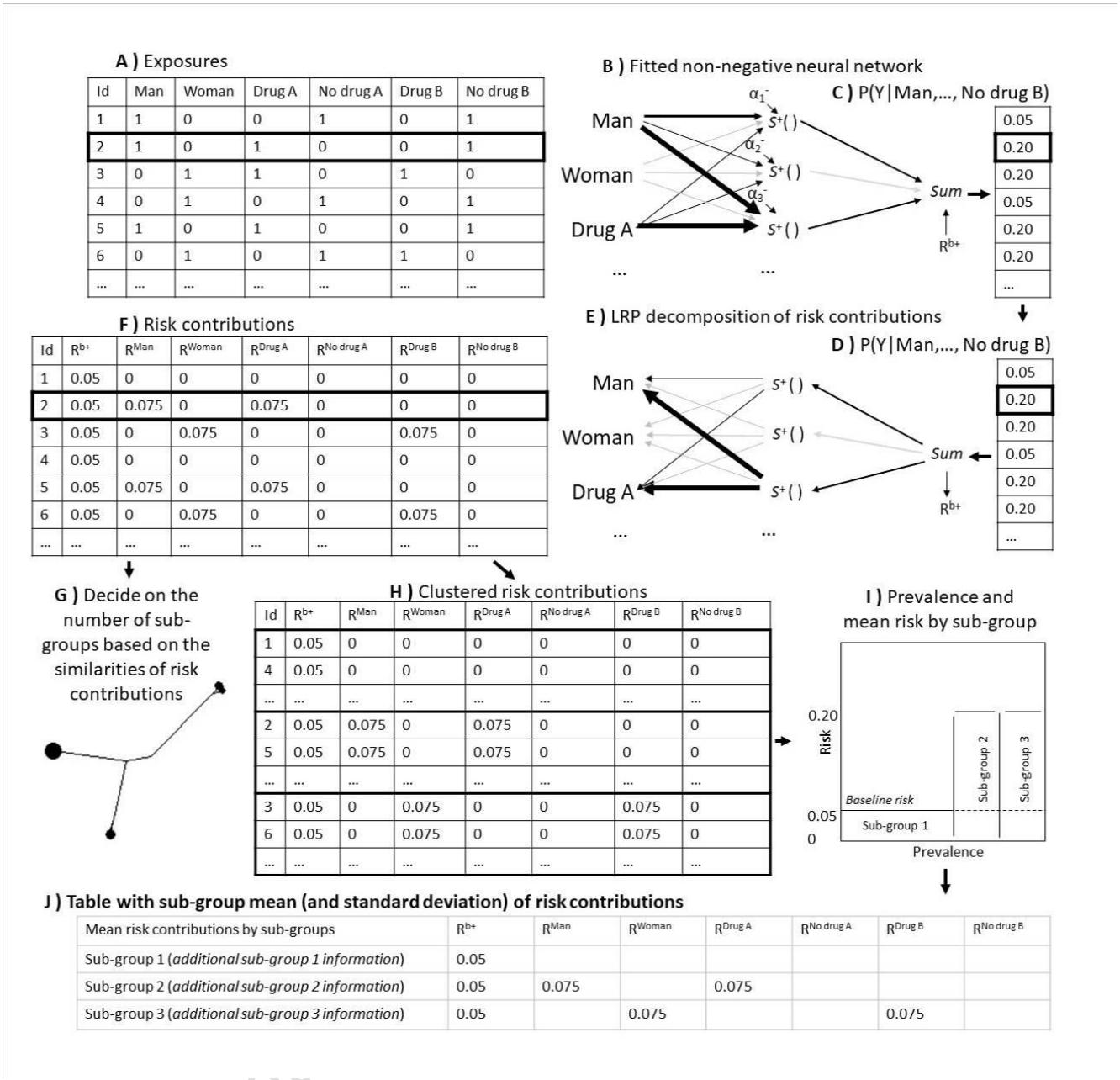
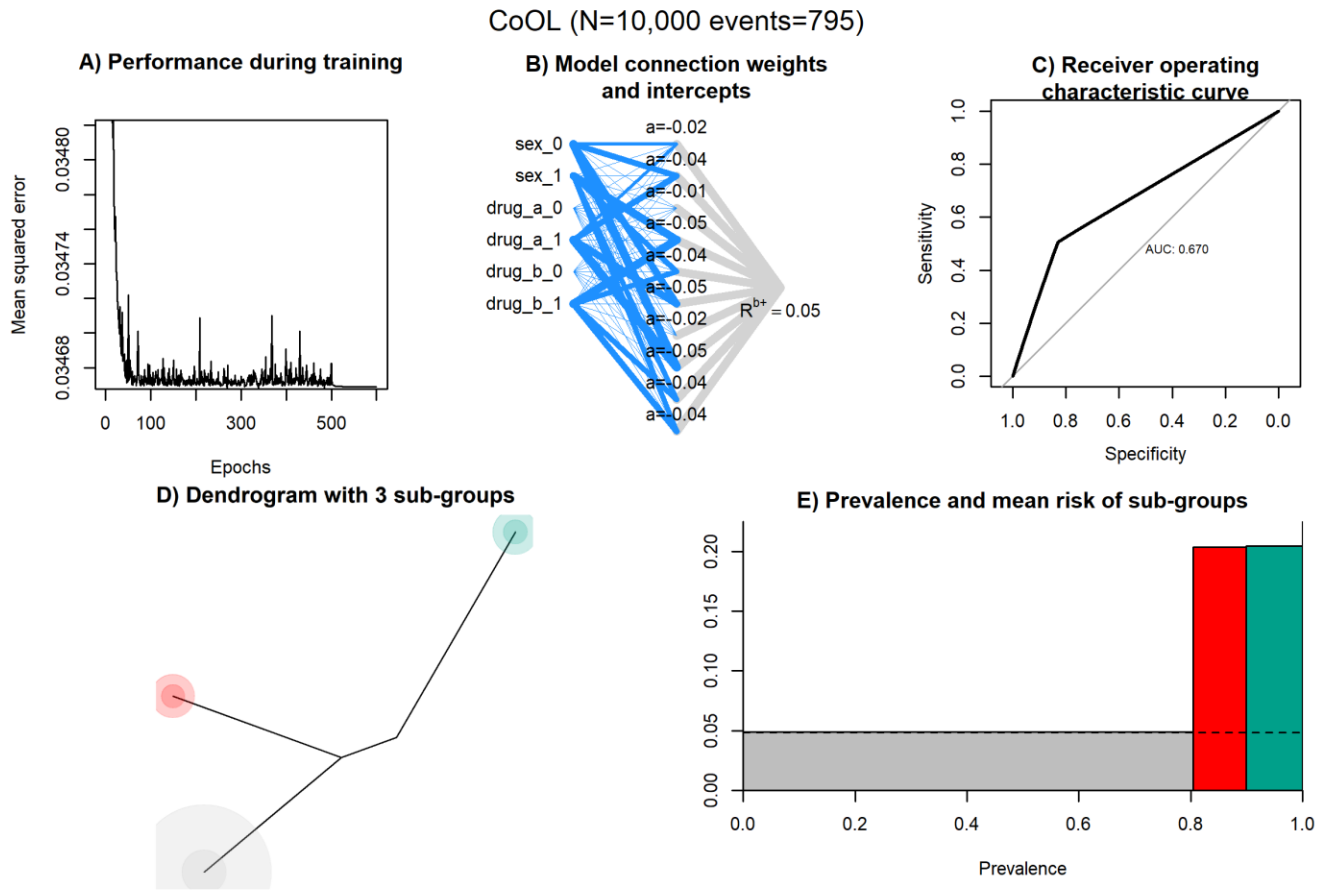


Figure 3. Workflow of the computational phase of CoOL

The flowchart of how sub-groups are identified as part of the computational phase of Causes of Outcome Learning. A) The expanded dataset of sex (one variable for man, one for woman), drug A (one variable for drug A, one for no drug A), and drug B (one variable for drug B, one for no drug B). B) The fitted non-negative model is illustrated. Wide edges indicate large connection parameters. C-D) The predicted risk, $P(Y|X)$. E) The predicted risk is decomposed using LRP to risk contributions of the baseline, R^{b+}, and exposures, R^{X}. F) The risk contribution matrix. G) A dendrogram to help decide on the number of sub-groups. H) Clustered risk contribution matrix into sub-groups. I) Prevalence and mean risk by sub-group plot. This plot indicate areas for greater public health impact. J) A table with sub-group mean of risk contributions. It can hold more information which can be useful when developing hypotheses, such as quantifications of the excess proportion of all cases found is this sub-groups when considering the prevalence of the subgroup, the risk in the sub-group and the baseline risk.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .



F) Mean risk contributions by sub-group
(Standard diviation)
[mean risk contribution if other exposures are set to 0]

	Baseline_risk	drug_b_1	drug_b_0	drug_a_1	drug_a_0	sex_1	sex_0
Sub-group 1: n=8038, e=393,Prev=80.4%, risk=4.9%, excess=0.4%, Obs risk=4.9% (4.4-5.4%) Risk based on the sum of individual effects =4.9%	4.9% (0.0%) [4.9%]						
Sub-group 2: n=950, e=194,Prev=9.5%, risk=20.3%, excess=18.5%, Obs risk=20.4% (17.9-23.2%) Risk based on the sum of individual effects =5.2%	4.9% (0.0%) [4.9%]	7.8% (0.1%) [0.4%]				7.4% (0.1%) [0.0%]	
Sub-group 3: n=1012, e=208,Prev=10.1%, risk=20.5%, excess=19.9%, Obs risk=20.6% (18.1-23.2%) Risk based on the sum of individual effects =4.9%	4.9% (0.0%) [4.9%]			7.7% (0.0%) [0.0%]			7.7% (0.0%) [0.0%]

Figure 4. Results of the computational phase of CoOL

THE MAIN RESULTS ARE COMBINED IN ONE PLOT. A) PERFORMANCE MEASURED BY THE MEAN SQUARED ERROR BY EPOCH. B) A VISUALISATION OF THE FITTED NON-NEGATIVE NEURAL NETWORK. THE WIDTH OF THE LINE INDICATE THE STRENGTH OF EACH CONNECTION. C) A PLOT ON PREDICTION PERFORMANCE AS MEASURED BY A ROC CURVE. D) A DENDROGRAM COLORED BY 3 GROUPS. E) THE MEAN RISK AND PREVALENCE BY SUB-GROUPS. F) THE TABLE WITH THE MAIN RESULTS FOR THE WORKING EXAMPLE. "N" IS THE TOTAL NUMBER OF INDIVIDUALS IN THE SUB-GROUP, "E" IS THE NUMBER OF EVENTS / INDIVIDUALS WITH THE OUTCOME IN THE SUBGROUP, "PREV" IS THE PREVALENCE OF THE SUB-GROUP, "RISK" IS THE MEAN RISK IN THE SUB-GROUP BASED ON THE MODEL, "EXCESS" IS THE EXCESS FRACTION BEING THE PROPORTION OUT OF ALL CASES WHICH ARE MORE THAN EXPECTED (MORE THAN THE BASELINE RISK) IN THIS SUB-GROUP, "OBS RISK" IS THE OBSERVED RISK IN THIS SUB-GROUP (95% CONFIDENCE INTERVAL IS CALCULATED USING THE WALD METHOD IN [74]), "RISK BASED ON THE SUM OF INDIVIDUAL EFFECTS" IS THE RISK SUMMED UP WHERE ALL OTHER EXPOSURES ARE SET TO ZERO. FOR THE 3 ESTIMATES PRESENTED AT EACH VARIABLE BY EACH SUB-GROUP, THE FIRST ESTIMATE IS THE MEAN RISK CONTRIBUTION, THE ESTIMATE IN PARENTHESES IS THE STANDARD DEVIATION, AND THE ESTIMATE IN BRACKETS IS THE RISK CONTRIBUTION HAD ALL OTHER EXPOSURES BEEN SET TO ZERO. THE BASELINE RISK IS BY DEFINITION THE SAME FOR ALL GROUPS.

medRxiv preprint doi: https://doi.org/10.1101/2020.12.10.20225243; this version posted December 11, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted medRxiv a license to display the preprint in perpetuity.

It is made available under a CC-BY-NC-ND 4.0 International license .

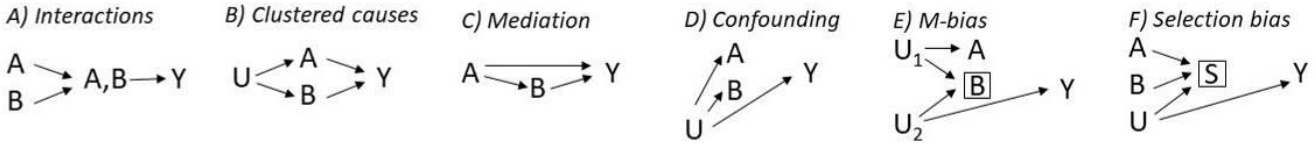


Figure 5. Six causal structures causing co-occurring associations

A and B denotes measured exposures of interest, U denotes an unmeasured cause of A and B, Y denotes the outcome, S denotes a selection mechanism. All six causal structures result in an increased co-occurrence of A and B in the Causes of Outcome Learning approach. It only applies for interactions that the combined effect is larger than the sum of the individual effects. A) Interaction - A and B jointly affect Y, and thus occur often together when assessing risk contributions (see also [31]). B) Clustered causes - A and B occur more often together due to U. C) Mediation - Since B is caused by A, A and B often occur together (see also [31]). D) Confounding - If U is a cause of A, B and Y, all variables occur often together (See also [6]). E) M-bias - Selection on B can cause a non-causal association between A and B, and A and Y (See also [39]). F) Selection bias - Conditioning on S creates a non-causal association between A, B and Y (See also [39]).

Pre-print (not peer-reviewed nor accepted by a journal)