JDSSV

Journal of Data Science, Statistics, and Visualisation

April 2024, Volume IV, Issue III. *doi: 10.52933/jdssv.v4i3.72*

RoA: Visual Analytics Support for Deconfounded Causal Inference in Observational Studies

Dennis Dingen
Eindhoven University of Technology

Marcel van 't Veer
Catharina Hospital Eindhoven
Eindhoven University of Technology

Tom Bakkes
Eindhoven University of Technology

Erik Korsten
Catharina Hospital Eindhoven

Arthur Bouwman
Catharina Hospital Eindhoven

Jarke J. van Wijk
Eindhoven University of Technology

Abstract

The gold standard in medical research to estimate the causal effect of a treatment is the Randomized Controlled Trial (RCT), but in many cases these are not feasible due to ethical, financial or practical issues. Observational studies are an alternative, but can easily lead to doubtful results, because of unbalanced selection bias and confounding. Moreover, RCTs often only apply to a specific subgroup and cannot readily be extrapolated. In response, we present Rod of Asclepius (RoA), a novel visual analytics method that integrates modern techniques designed for identification of causal effects and effect size estimation with subgroup analysis. The result is an interactive display designed to combine exploratory analysis with a robust set of techniques, including causal do-calculus, propensity score weighting, and effect estimation. It enables analysts to conduct observational studies in an exploratory, yet robust way. This is demonstrated by means of a use case involving patients undergoing surgery, for which we collaborated closely with clinical researchers.

Keywords: visual analytics, causal inference, confounding, observational study, exploratory data analysis.

1. Introduction

The process of determining and estimating the relationship between a cause and effect is referred to as causal inference. This process is intrinsically important in scientific disciplines and especially when conducting observational studies to analyze data that have already been recorded. In these types of studies, the development of prediction models can benefit from the theory behind causal inference to reduce or minimize the unfavorable effect that confounding variables have on the estimation of a cause and effect relationship (Pearl et al. 2016a). Because no control on the data acquisition can be exerted (anymore), the researcher is required to have a sufficient understanding of how the causal relationship of interest is embedded in the bigger causal structure to properly apply the theory. Such an understanding is often expressed using a graph that maps variables onto nodes and causal relationships onto directed edges (arrows). The graph structure is obtained by synthesizing scientifically established results with empirical domain knowledge and educated guesses, which is therefore inherently prone to different interpretations and usually leaves room for debate among domain experts (Hill 1965).

Several statistical techniques exist that minimize confounding effects to assess differences in groups, such as matching, weighting or stratification (Leite 2016a). Traditional software packages like SPSS (IBM Corp. 2021) and interactive notebooks based on R (R Core Team 2021) or Python (Python Software Foundation 2021) currently offer only part of an interactive process to facilitate these techniques (see Table 4 in Appendix B). Some interactive visualization methods have been developed in support of causal inference, but generally these are limited to causal graph discovery and lack more thorough support for deconfounded effect estimation.

The goal of our work is to directly address practical aspects of causal inference by means of a novel visual analytics tool, dubbed Rod of Asclepius (RoA). To this end, we utilize statistical methods interactively to support causal effect estimation for observational studies in an exploratory setting. This enables researchers to efficiently explore the data and investigate implications of a causal model under consideration during debates. The gathered insights can be used to form new hypotheses and support confirmatory studies later on. Although our tool is generally applicable, it was developed within clinical setting as part of collaboration between academia, industry and healthcare providers (e/MTIC 2021).

2. Method

This section starts with the introduction of two different study designs used for causal inference; the randomized controlled trial and the observational study. Next, the workflow for estimating causal effects in an observational study is discussed. A key step

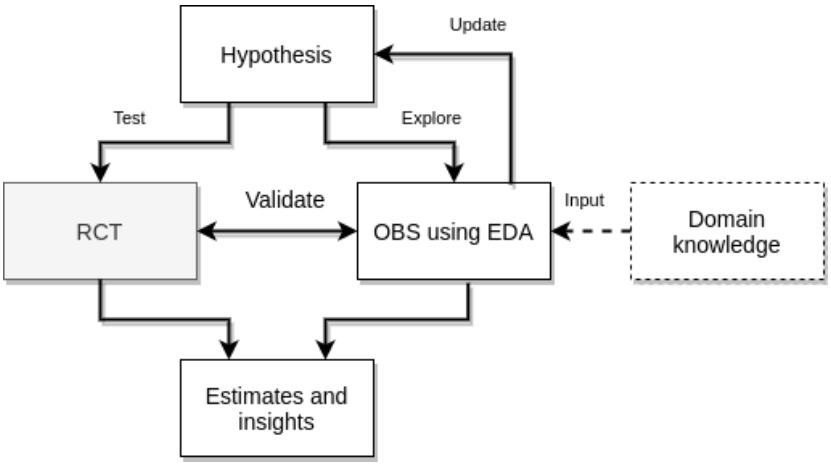


Figure 1: The operational context of the Randomized Controlled Trial (RCT) and the Observational Study (OBS) augmented with Exploratory Data Analysis (EDA).

of this workflow is controlling for confounding, by balancing a proper set of variables associated with the causal relation being studied. In order to find such a set, we need to analyze the structure of the causal graph that embeds the causal relation being studied by utilizing a method developed by Pearl, called "do-calculus" (Pearl et al. 2016b). Finally, we detail the translation of these steps and methods into interactive visualizations and how these have been integrated into the design of our visual analytics system to support observational studies in an exploratory setting.

2.1. Causal inference

The gold standard for estimating the causal effect size of a specific phenomenon onto another phenomenon is the Randomized Controlled Trial (RCT)(Beal and Kupzyk 2014; Fonarow 2016). In a RCT, the *a priori* control of variables is used by the researcher to randomly assign subjects to (treatment) groups. If applied adequately, this avoids confounding effects caused by other associated variables by assuring that the groups under consideration are as similar as possible in statistical terms. It is, however, not always possible or responsible to conduct a RCT to properly estimate a causal effect, due to practical or ethical considerations. In those cases, researchers can fall back to conducting an Observational Study (OBS).

The main characteristic of observational studies is that no variable can be manipulated beforehand and therefore a researcher can only retrospectively inspect the recorded data. With no a priori control over variables, the group membership (treatment) of the subjects being studied can therefore not be influenced. The major consequence is that confounding variables can affect an estimation of the treatment effect. Hence, the influence of confounding variables needs to be adjusted for when conducting an OBS to obtain sound causal effect estimates, which is done using domain knowledge (Collins et al. 2020). Our visual analytics tool is designed to support the operational context for conducting an OBS in such a way that domain knowledge can be easily integrated within an exploratory setting. This also enables us to conduct an OBS more easily in a complementary setting with respect to the RCT, as shown in Figure 1.

2.2. Estimating treatment effects in an observational study

To estimate the causal effect size of one phenomenon Z on another Y, for a given causal relationship $Z \rightarrow Y$, we follow the analysis workflow shown in Figure 2. The goal is to obtain unconfounded treatment effect estimates. After selecting a causal relationship, we inspect the remaining covariates for their potential influence on the treatment effect estimation. For this, we need to distinguish between the following:

1. The (potential) size of the influence a variable can have across treatment groups onto another variable.
2. Whether or not it is confounding the treatment effect estimate (i.e., actually exerting its influence).

Firstly, the (potential) size of the influence of a variable depends on the similarity of its distributions across the treatment groups. Thus, if these distributions are similar or made similar by "balancing" them, the potential influence is negligible. This concept is illustrated in Figure 3a. In addition to visual inspection, we need to quantify whether or not its distributions are similar statistically. For this purpose, we use the Standardized Mean Difference (SMD).

Secondly, whether a covariate can confound the treatment effect estimate depends on its position in the causal diagram. For example, a confounding variable can directly influence the causal relation under consideration, as illustrated in Figure 3b. In general, however, the situation may be more complex. The causal diagram is defined either: by experts, by an automated mining algorithm, or a combination of both. Hence, in order to obtain unconfounded treatment effect estimates, we need to make sure that the influence of all confounding covariates is eliminated. To this end, we use propensity score weighting to balance dissimilar distributions of covariates. The set of covariates to balance can be chosen manually or determined by the application of do-calculus based on the causal diagram. These steps are indicated in the top part of Figure 2. When the influence of confounding covariates is eliminated, unconfounded treatment effect estimation is possible, which is the last step of the workflow.

Furthermore, the workflow is adopted specifically to facilitate interactive exploratory data analysis of generic datasets of common size in a clinic in the medical field; dozens of variables and thousands of subjects. On top of that, propensity score weighting does not impose restrictions on the type of model used for treatment effect estimation. Generally,

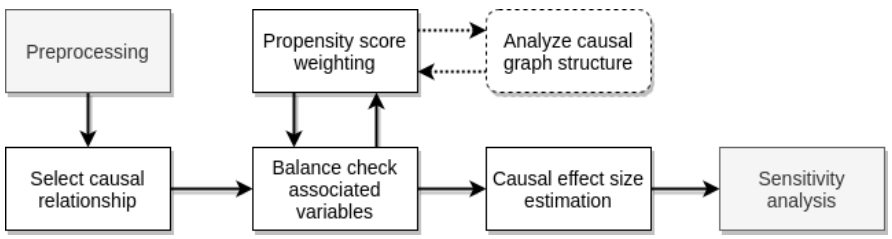
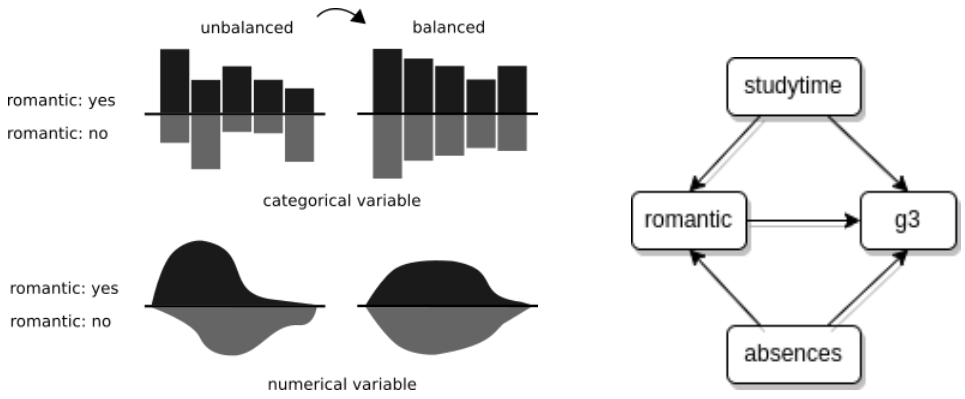


Figure 2: Main analysis workflow based on (Leite 2016b) with additions for clarification and integrative purposes. The data preprocessing step is considered outside of our scope and the sensitivity analysis step is an extra step to be considered in future work.



(a) (b)

Figure 3: (a) Illustrations of unbalanced and balanced distributions for a categorical and numerical variable. (b) The variables *studytime* and *absenses* acting as confounding variables on the causal relationship between *romantic* and *g3*.

one should in principle be able to use alternative statistical techniques without affecting the workflow nor the design of RoA too much. To further illustrate the steps of the workflow, we consider a running example based on a publically available dataset about final grades of math students (Cortez and Silva 2008). For mathematical details behind the analysis workflow for causal inference in an OBS, shown in Figure 2, see Appendix A.

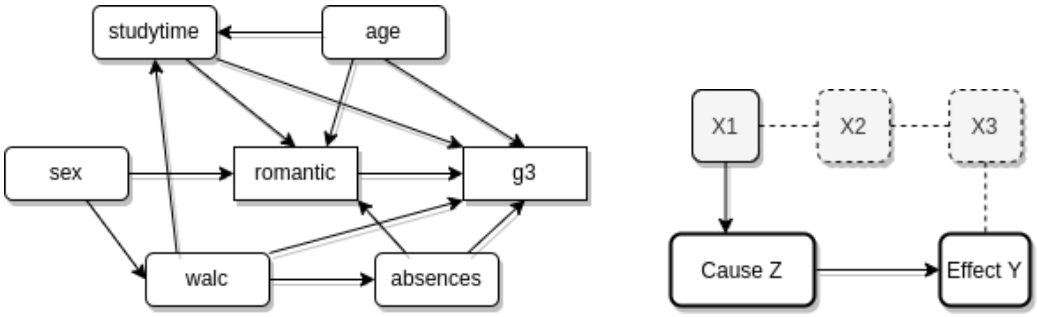
Select causal relationship

Suppose we want to estimate the effect size of having a romantic relationship on the (third and) final grade of math students for a math exam: *romantic* $\rightarrow$ *g3*, where *romantic* and *g3* are a dichotomous (binary) and real-valued variable, respectively. Suppose also that the data had already been recorded beforehand, and we could not have randomly assigned romantic relationships to students for our study. Given this situation, we conduct an observational study to estimate the effect size of having a relationship on the final math grade of students. If we knew beforehand, with absolute certainty, that no confounding variable existed, we could immediately estimate the causal effect, but generally, this is not the case, and we need to proceed with the intermediate steps of the workflow.

Balance check of associated variables

In order to obtain a sound unconfounded effect estimate, we need to ensure that variables that could have had a confounding influence on our causal relationship of interest are balanced. A balanced variable has statistically similar distributions in the treatment groups, which can be quantified with the SMD. The SMD is computed differently for continuous and dichotomous variables (see Equation 9 and Equation 10 in Appendix A), but should in both cases be (close to) zero for optimal balance. Hence, in practice, the distributions are considered similar when the SMD is within an acceptable range. Commonly used cut-off values for the acceptable range are $[-0.1, 0.1]$ (Austin 2011) and $[-0.25, 0.25]$ (Stuart and Rubin 2007; Stuart 2010).

The SMD values for our covariates can also be inspected visually using a decorated



(a) (b)

Figure 4: (a) Causal diagram for estimating the effect of having a romantic relationship on the final math grade *g3*. (b) A backdoor path as a generalization of the basic confounding variable. It is a sequence of variables X that connects the cause variable Z with the effect variable Y, whereby the first variable $X1$ of the sequence is connected via an arrow directed towards the cause Z.

love plot, as shown in Figure 5a. The plot is sorted on decreasing SMD value for the variables. The acceptable SMD range is indicated with dotted lines centering around the zero axis and the tickmarks indicate which variables are part of the adjustment set that is used for the balancing process, which is discussed below. Please note that while the SMD indicates the potential influence of a covariate, it does not imply whether or not a covariate acts as a confounding variable for our causal relation of interest. As stated before, this depends on the position of the covariate in the causal graph.

In a properly conducted RCT, the randomized assignment of treatments to subjects to create the treatment and control groups ensures that the SMD values for all covariates are close (enough) to zero. Consequently, the influence of all possible confounding covariates is eliminated effectively at once. In that case, the positions of covariates in the causal diagram are no longer of interest. When conducting an OBS, we do not have the benefits of randomization, and we need to ensure all relevant covariates are balanced. The main problem in doing so is that covariates act together, depending on the structure of the causal diagram. As a result, an improperly picked adjustment set of covariates to balance can lead to even more confounding and bias in the effect estimates. Therefore, the strategy of adding covariates to the adjustment set based on only SMD values is expected not to work. Hence, current (clinical) guidelines often suggest adding a covariate based on domain knowledge and whether it is associated with the causal relation of interest (Beal and Kupzyk 2014; VanderWeele 2019; Witte and Didelez 2019; Loh and Vansteelandt 2020; Talbot et al. 2021).

Propensity score weighting and causal graph analysis

As mentioned before, the proper set of variables to balance depends on the causal graph structure. This is where do-calculus helps us out. For our running example, we assume that the bigger causal graph embedding our relationship of interest is as shown in Figure 4a. In practice, a graph like this would be delineated during expert discussions, optionally with the aid of graph mining algorithms (see Table 5 in Appendix B).

Following do-calculus, adjusting for the effects of confounding variables entails the elim-

ination of all so-called "backdoor paths" in the causal graph in which the relationship of interest is embedded. A backdoor path resembles a generalized confounder; it is a substructure consisting of a sequence of one of more variables that form a path from the cause to the effect. The first variable on this path needs to point to the cause, but there are no further constraints on the directions of the remaining edges (arrows) on the path, as shown in Figure 4b. For example, $romantic \leftarrow studytime \leftarrow age \rightarrow g3$, $romantic \leftarrow sex \leftarrow walc \rightarrow g3$ and $romantic \leftarrow sex \leftarrow walc \rightarrow absences \rightarrow g3$ are backdoor paths in the graph. The variable *walc* indicates "weekend alcohol consumption".

Applying do-calculus yields possible sets of variables we can balance to block confounding effects via backdoor paths. In our example, out of all possible adjustment sets yielded by do-calculus, we have picked the set $\{absences, age, sex, studytime\}$. If do-calculus does not yield a possible adjustment set, we at least know which parts of the causal graph are problematic. This information is helpful during the next round of debate or for determining new real-world measurements, which may lead to an updated graph. Alternatively, we can accept that our estimates are confounded and try to reduce confounding manually as far as possible by falling back to using empirical rules in conjunction with our causal diagram (Leite 2016b).

After deciding on the adjustment set, we need to balance its covariates. Generally, a variable can be balanced by using matching, weighting, stratification, or randomization. The use of matching is an option when conducting an OBS if enough subjects are available in the dataset, and one knows which variables to match. We have chosen to adopt the weighting (reshaping) of the distributions based on propensity scores because it is often more readily applicable in practice and easily integrates into our analysis workflow. Furthermore, the method is computationally efficient and does not impose assumptions on our causal relation, such as linearity, and therefore leaves room for choice of the effect size model later on. The use of stratification is an option when separate models for the subgroups (strata) are of interest.

The propensity score for a given subject is defined as the conditional probability of treatment assignment based on the values of covariates (in the adjustment set) and is commonly computed using logistic regression (see Equations 3–5 in Appendix A). Once computed, it is used as a weight for computing the balanced distributions (Figure 3a) to obtain adjusted SMD values (Figure 5a). In the latter figure, our adjustment set $\{absences, age, sex, studytime\}$ is used, as indicated by the tickmarks. The SMD values before balancing are indicated with dashed circles connected with a dotted curve, while the values after balancing are indicated with solid circles connected with a solid curve (these types of curves resemble snakes curling around the Rod of Asclepius (RoA) and lead to the name of our tool). The unbalanced and balanced SMD values summarize differences between pairs of distributions shown in mirrored-liked fashion in Figure 3a, on the left and right sides, respectively. Please note that because covariates act as a whole depending on the causal graph, balancing a particular covariate might cause a shift in the balanced SMD value of other covariates.

Finally, how the propensity score is used as a weight depends on the estimand one wants to use for the treatment effect estimation (see Table 3 in Appendix A). We use the Average Treatment Effect (ATE) and the Average Treatment Effect on the Treated (ATT). The ATE estimates the treatment's effect on the entire population, while the

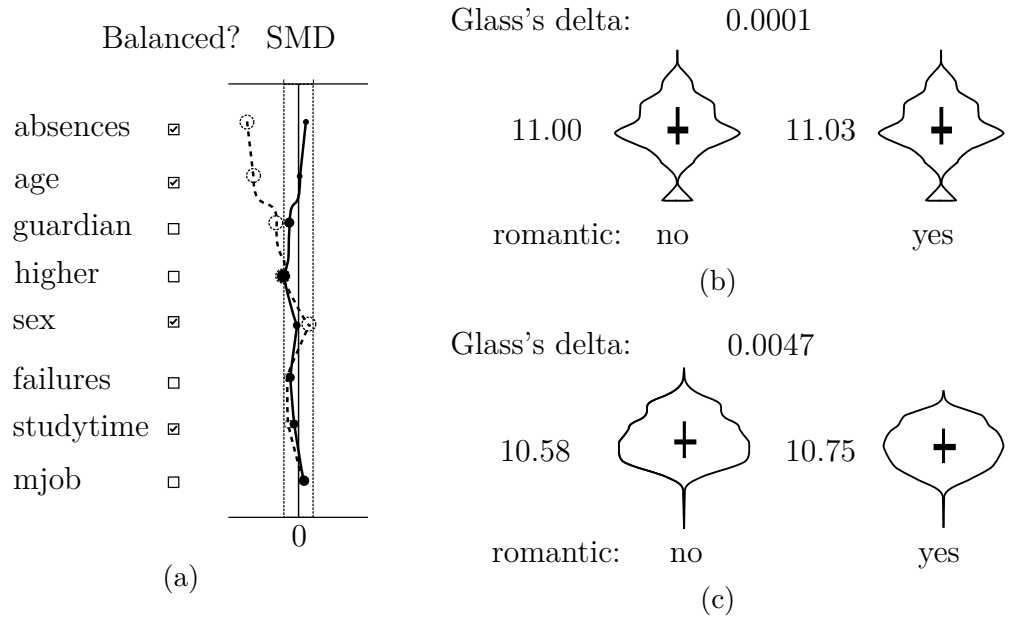


Figure 5: (a) Similarity of distributions of potentially confounding variables across the levels of the variable *romantic*. Open circles and connected with a dashed curve indicate SMD values before balancing, while filled circles connected with a solid curve indicate the SMD values after balancing some of the variables (those indicated by the check marks). (b) Effect estimation of having a romantic relationship on the final math grade before balancing. The distributions of math scores are shown using violin plots. (c) Same estimation as (b) but after balancing.

ATT estimates the effect the treatment has on only the treated group (see Equations 1–2 and Equations 7–8 in Appendix A). After propensity score weighting, the treatment effect can be estimated, but as discussed below, we can improve the estimate by using an additional step.

Causal effect size estimation

To estimate the effect size of having a romantic relationship on the final math grade, we subtract the (weighted) mean final grade *g3* of the group having no romantic relationship from the (weighted) mean final grade of the group having a romantic relationship (see Equation 11 in Appendix A). The weights used are computed in accordance with the required estimand; the ATE or ATT. For improved treatment effect estimation, we adopt the *doubly robust method* that substitutes the expected value for individual subjects with a predicted expected value based on the propensity score (see Equations 12–13 in Appendix A).

Next, we compute whether this difference is statistically significant. In Figure 5b and Figure 5c, the effect estimates are shown without and with the prior balancing of the adjustment set. There is, however, also an alternative estimation method called Glass's Delta (see Equation 14 in Appendix A) that uses a form of normalization, which is shown on the top of the figures. The motivation for this is that the raw mean difference is generally not stable and homogeneous because it depends on the unit of measurement.

Without balancing the covariates in the adjustment set, we observe a statistically non-significant slight difference of 0.03 in the mean math grades with a Glass's Delta of 0.0001. Next, after balancing the covariates in the adjustment set, we observe a non-significant yet slightly bigger difference of 0.17 in the mean math grades with a Glass's Delta of 0.0047. Therefore, conditional on the correctness of our causal graph and the set of covariates to be balanced yielded by it, we can conclude that having a romantic relationship caused a slight, yet statistically insignificant, increase of 0.17 in final math grade for the students involved in a romantic relationship.

2.3. Visual analytics solution

In this section, we discuss the start of the art in visual analytics for causal inference. Next, we detail the design of our tool RoA by showing how the parts discussed in Secion 2 have been integrated via interactive visualizations.

State of the art

Although visualization for causality analysis is not new, only relatively few papers have been published on the topic. Recent papers typically deal with one or more of the subareas of causal inference (Tikka and Karvanen 2017):

1. the discovery of the causal model (from data);
2. the identification of causal effects using a known model;
3. the actual estimation of an identified causal effect from data.

However, most papers are related because they focus on visualizing (some type of) causal relations. Others are related more strongly because they are centered around the theory developed by Pearl and others, for either Bayesian networks or causal diagrams.

Earlier visualization methods focused on showing causal relationships using connected polygonal shapes (Elmqvist and Tsigas 2003) and animated graphs (Kadaba et al. 2007). Later, a tool was developed for causality analysis in biological pathways using a combination of node-link diagrams, arc diagrams, and animation (Dang et al. 2015). Two systems explicitly designed for causal inference in the context of event sequence data have been published by (Xie et al. 2021) and (Jin et al. 2021). A user study on medication recognition was published by Yen et al. (2019).

More recent work has been primarily focused on the first subarea in causal inference– the discovery of causal models. Visualization support for Bayesian network structure learning was published by Vogogias et al. (2018). It was shown that drawing all the edges for just a couple of networks simultaneously connected to a small number of shared nodes made the visualization already hard to read. The solution involved augmenting adjacency matrices with glyphs comprised of one or two colored triangles to encode an edge's networks and direction. Many hundreds of networks could be compared by showing multiple augmented matrices side-by-side after applying a form of graph filtering on nodes and edges.

Next, we have two systems published by (Wang and Mueller 2016, 2017). Both utilize the theory of causal inference developed by Pearl and others. The first one, called *The*

Visual Causality Analyst, offers an interactive interface for causal reasoning. The user is enabled to draw a causal diagram in the form of a graph. The system then automatically computes a regression model for each node, such that the variable associated with the node is the responder, and the variables in the parent nodes are the covariates. The model weights are then used to indicate the strength of the relationships on the associated edges. Notably, a novel method was proposed to transform categorical variables into numerical ones to aid the computations. The second system is designed for causal graph mining. It shows associations between variables for data partitions based on which candidate causal graph can be computed. The associations are shown in a parallel coordinates plot and heatmap, respectively. These computed graphs are aggregated using counts or strengths of mined relations into a single graph representation to show the user's overall causal patterns using curved edges.

The first system shows the graph with color-coded nodes and edges, with an arrowhead on the center of each edge indicating the direction of the relation. The color of a node encodes the variable type (blue for numerical and yellow for categorical). The color of an edge encodes the type of causal relation (green for positive relations, red for hostile relations, and yellow if the source or target node is associated with a categorical variable). This way, the system can explore the causal graph's properties while filtering out edges using a strength threshold. The second system comes with an improved visual encoding for the graph. Nodes have a blue rectangle for categorical variables and yellow for numerical ones, with varying width, depending on the goodness of fit measure of the underlying regression model. The width of an edge corresponds to the strength of a relationship, computed with regression coefficients, and its color is green for positive relations and red for negative ones. A yellow edge reflects a compound relationship between levels of a categorical variable and other variables. Next to the edge, a red minus sign or green plus sign is shown when a relationship should be added or removed from the aggregated model. The resulting causal graph is shown as a path diagram, such that nodes are more aligned and the overlapping of the (now curved) edges seems to be reduced. A plus or minus sign next to the edge is used to communicate whether an edge should be present or not in the mined aggregated graph.

Finally, we have a system designed by (Xie et al. 2020). The system offers the user an interactive graph visualization with the option to collapse subgraphs in a node. Additional panels show the histograms of all variables in the dataset along with more detailed information about all the values. The *F-GES* algorithm is applied interactively for conditional independence testing of variables. The resulting value is encoded as the thickness of the edges connecting nodes representing the tested variables. The user can then fix values for specific variables or change the attribution of variables in the causal discovery process to explore what-if scenarios.

In conclusion, the systems designed by Wang and Mueller (2016, 2017) and Xie et al. (2020) are the most closely related to our work. These systems comprised the first and third subareas of causal inference and were not designed for explicitly adjusting for confounding. Furthermore, regression models assume that treatment (nodes) and effect (nodes) are related linearly. In our work, we address these aspects in our work to accommodate for observational studies and focus while focusing more on the second and third subareas of causal inference.

Visualizing causal graphs

An important aspect of the systems discussed in the previous section is the visual encoding used for the causal graphs. Nodes are colored depending on the variable type, and the directed edges (arrows) depend on the combination of connected variable types. An edge's width is used to reflect a relationship's strength, corresponding to regression model coefficients. Small icons can be used near edges or nodes to convey additional information as well. More details on on-edge encoding methods have been published Nobre et al. (2019).

Another important aspect is path analysis, for which dedicated support should be considered. There is evidence that a node-link diagram outperforms matrix-based visualizations for path analysis (Ghoniem et al. 2004). Topological sorting of the graph is helpful, but a sequential layout did show a significant difference in understanding indirect causal relationships, according to Bae et al. (2017a). A node's position and the number of connected edges were critical visual cues for finding root causes and derived effects. The graph also displayed strength and certainty information encoded through edge width and color brightness. Also, hierarchical graph structures performed better than energy-directed ones.

For visualization, tapered edges or arrows are well suited for causality, and one should consider the width of arrows over hue for strength. In another paper, Bae et al. (2017b) presented findings of visual representations of cause and effect relationships regarding their direction (using arrows and tapered lines), strength (using hue, width, and numeric values), and uncertainty (using granularity fuzziness and numeric values). The authors concluded that:

- arrows and tapered lines both work well,
- width is preferred over hue for encoding strength, and,
- brightness or fuzziness are preferred over granularity for encoding certainty.

Additionally, Guo et al. (2015) recommends using brightness, fuzziness, and grain to depict causality clues, but the design should be carefully investigated for the task at hand.

Design of RoA

Most papers seem to focus on the first and third subarea of causal inference. Generally, we found that if authors mentioned the problem of confounding, it was usually not adequately adjusted for in the estimates. Furthermore, typically, the use of a method such as regression to estimate causal effects demand assumptions like linearity, which is not necessarily adequate. We address these aspects in our work to accommodate for observational studies by focusing more strongly on the second and third subareas of causal inference. In particular, our solution exhibits the following properties:

- a single cause and effect relationship of interest is considered at a time;
- no assumption is made about the type of causal relationship (in the design of our system as a whole);

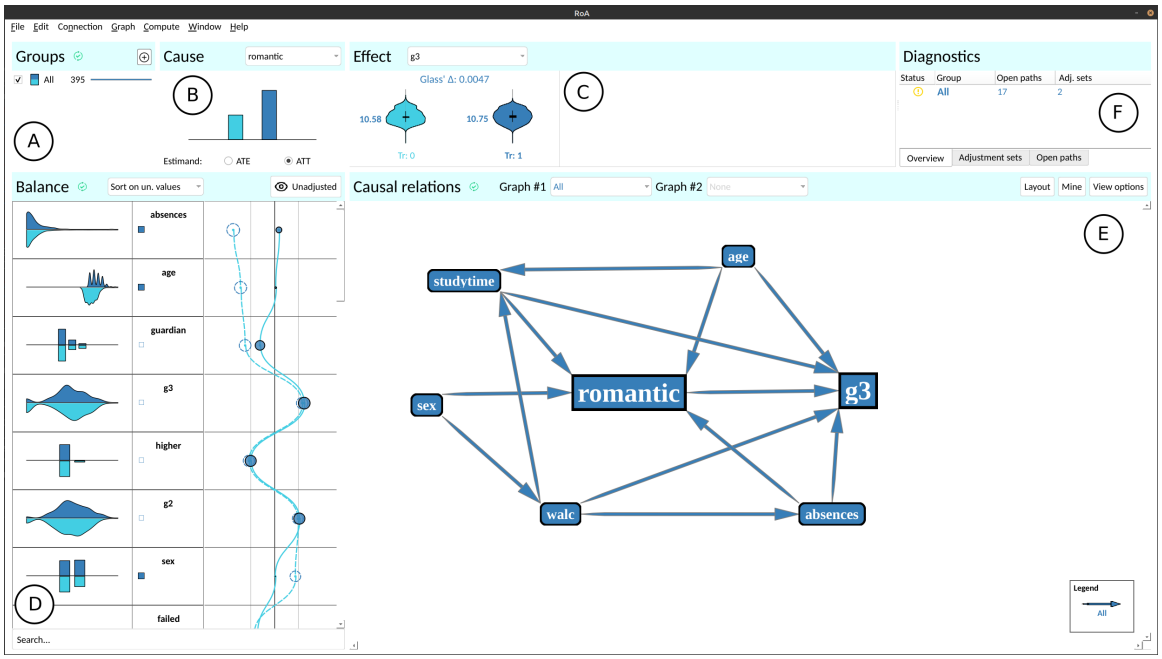


Figure 6: Overview RoA. (a) Groups and subgroups panel. (b) Cause variable selection panel. (c) Effect variable selection panel. (d) Variable balance panel. (e) Causal graph editor. (f) Causal graph diagnostics panel.

- alternative methods for techniques like matching and stratification for ensuring balance could be used instead of weighting;
- alternative causal effect estimation techniques can used, like the doubly robust method based on regression (Leite 2016c);
- manual selection of a set of variables to adjust for confounding is supported;
- optimal adjustment sets are computed, if possible, based on the causal graph;
- interactive subgroup analysis for subgroups is supported; and
- automated graph mining algorithms have been implemented to aid expert discussions.

A screenshot of RoA is presented in Figure 6. The key motivation for developing RoA was to provide support for real-time interactive exploratory analysis during expert discussions. To this end, RoA offers interactive design and diagnostics of the causal graph diagram, along with dynamic balancing of either computed or manually selected sets of confounding variables and automatic updates of the causal effect estimations.

For designing RoA we listed a number of design requirements (see Table 1) based on the analysis workflow discussed in Section 2.2. We have mapped these requirements into visualization requirements for the tool, for which acronyms are listed in the third column. These are used in the remainder of this section to indicate which part of RoA fulfills each requirement. For convenience, the fourth column lists the corresponding panel labels in Figure 6. The algorithms and packages used to implement the computational aspects of the tasks are listed in Table 2.

Table 1: The mapping of the main consecutive workflow tasks onto visualization requirements: (a) Selection of a causal relationship. (b) Balance check of associated variables. (c) Analysis of the causal graph structure. (d) Causal effect size estimation.

Tasks	**Visualization requirement**	**Acronym**	**Label in Figure 6**
Define (sub)groups	Show group names, associated colors and sizes	VR1	A
Select dichotomous cause (treatment) variable	Selector for variable showing distribution	VR2	B
Select continuous effect variable	Selector for variable showing distribution	VR3	C

(a) Selection of a causal relationship.

Tasks	**Visualization requirement**	**Acronym**	**Label in Figure 6**
Determine method for weighting and estimation (ATE / ATT)	Selector for weighting and estimation method	VR4	B
Compare unweighted and weighted distributions of variables for each level of the cause variable	Show both distributions	VR5	D
Inspect SMD values for the distributions of variables	Show SMD differences w.r.t. a treshold	VR6	D

(b) Balance check of associated variables.

Tasks	**Visualization requirement**	**Acronym**	**Label in Figure 6**
Mine causal graphs	Menu options	VR7	E
Edit causal graphs	Graph editor	VR8	E
Minimize visual complexity for graphs	Application of graph layout algorithm	VR9	E
	Removal of less important parts of graphs	VR10	E
Compare causal graph with mined version or of another group	Difference visualization for causal graphs	VR11	E
Analyze backdoor paths for a causal graph	List backdoor paths	VR12	F
	Highlight backdoor paths in the causal graph	VR13	F
Compute adjustment sets for a causal graph	Show adjustment sets (per group)	VR14	F
Import / export causal graphs	Menu options	VR15	-

(c) Analysis of the causal graph structure.

Tasks	**Visualization requirement**	**Acronym**	**Label in Figure 6**
Inspect effect variable across cause variable levels	Show effect variable distributions	VR16	C
Compute Glass's Delta	Show Glass's Delta	VR17	C

(d) Causal effect size estimation.

Before discussing the individual panels we start with the high-level functionality. Based on the main analysis workflow, shown in Figure 2, we have designed a more detailed operational workflow for RoA, shown in Figure 7.

Starting from an initial hypothesis, the researcher selects the causal relationship of interest. In response, RoA automatically computes balance measures for all remaining variables. In parallel, the researcher is presented with a minimal causal graph reflecting the causal relationship being studied. The researcher can then instruct RoA to automatically mine a bigger causal graph based on the data and/or make adjustments to the causal graph manually based on experts discussions.

When the causal graph has been changed, RoA computes backdoor paths and possible sets of variables to be balanced – adjustment sets – for optimal causal effect estimation, which can be selected by the researcher. If the set of variables to be balanced is changed, the causal effect estimates are updated automatically. Alternatively, the researcher can decide to define subgroups based on variables that are suspected of confounding the causal effect estimates. The consequence is of course that the causal estimates are now being studied for subgroups instead of the main group.

The researcher interacts with RoA through a series of panels, shown in Figure 6. Using panel (A), the researcher can define multiple subgroups to be used in the tool (in this example there is only one). The checkbox in front of a (sub)group indicates whether the

Table 2: Methods and packages used by RoA for computation.

Method	Package	Computational purpose
density()	stats (R Core Team 2021)	(Un)weighted density plots
hist()	graphics (R Core Team 2021)	Unweighted histograms
weighted.hist()	plotrix (Lemon 2006)	Weighted histograms
glm()	stats (R Core Team 2021)	Propensity scores, using logistic regression
Equations 7-8	Custom implementation*	Weights for ATE and ATT estimands
bal.tab()	cobalt (Greifer 2022)	(Un)weighted standardized mean differences
adjustmentSets()	dagitty (Textor et al. 2016)	Adjustment set based on causal graph
Equations 11-13 svyglm(), predict()	Custom implementation* survey(Lumley 2020)	Doubly robust effect estimation using propensity score weighting
pc.skel() and pc.or()	MXM (Lagani et al. 2017)	Causal graph mining
Orthogonal, Sugiyama, Energy	ogdf (Chimani et al. 2014)	Causal graph layout
markovBlanket()	dagitty (Textor et al. 2016)	Markov Blanket
Equation 14	Custom implementation*	Glass's Delta estimation

* See Appendix A.

group is currently enabled. The size of the group is shown on the right, along with a bar to show the relative sizes of subgroups. Each subgroup has also two associated colors (VR1). The hue of these colors (blue, red, ...) indicates the subgroup, and the intensity (light - dark) is used to distinguish between the two levels of the binary cause variable Z, which is selected in panel (B) by the researcher (VR2). In this panel also, the type of estimand is chosen (ATE or ATT), which is used for propensity score weighting and effect size estimates (VR4). We picked the ATT for now. The two colors associated with the group "All" are used to differentiate between levels of the treatment variable *romantic*.

Next, the effect variable Y is selected in panel (C) to convey our relationship of interest $Z \rightarrow Y$ (VR3). In this case, we see that the relationship *romantic* $\rightarrow$ $g3$ has been selected, as discussed earlier in Section 2.2. Violin-whisker plots show the distributions with the average values shown next to the center and Glass's delta is shown on top. If the two distributions are statistically different, the Glass's delta value is shown with a red background.

After selecting the causal relationship, we can inspect the balance of the other variables (covariates) with respect to the variable *romantic* by inspecting the SMD plot and distributions drawn next to them in panel (D). The distributions on the left correspond to subgroups defined by the cause variable, indicated with the two colors associated with the group "All" (VR5). Distributions of continuous variables are plotted using density plots and categorical variables using histograms. By plotting the two distributions of a variable in a "mirrored" fashion, visual comparison becomes more easy. In the second column, the variable name is shown along with checkboxes that indicate whether this

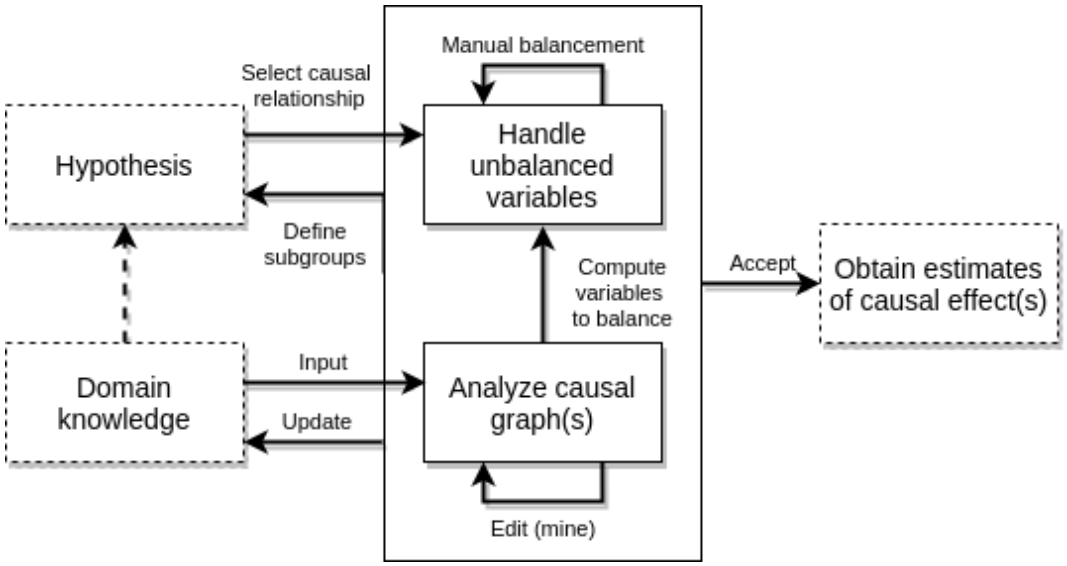


Figure 7: Workflow supported by RoA.

variable is in the adjustment set for balancing. This is indicated for each active group, but in this case only the group "All" is active. Clicking on a checkbox the researcher can manually include or exclude a variable to or from the adjustment set per group.

Furthermore, the SMD values are plotted on the right to indicate the difference between the distributions shown on the left (VR6). SMD values are connected using a curve to aid visual inspection. The button on the top right enables or disables showing of the unadjusted SMD values, which are drawn with dotted circles and dashed curves. In this case, the group colors are used to further differentiate the curves visually. On top of the panel is a selector to pick one of the sorting methods based on: variable name, unadjusted SMD values or adjusted SMD values.

The researchers involved in the study can draw their causal graphs in panel (E) (VR8). After selecting the causal relation using panels (B) and (C), an initial causal graph is generated automatically that reflects the cause and effect relationship. In this case we have $romantic \rightarrow g3$. In the graph, these variables are represented by rectangular boxes to emphasize the central role of the relationship. The researcher can now import or manually add more parts of the graph (VR15). In Figure 6, five more variables have been added along with some edges.

When the causal graph is edited, RoA automatically updates the diagnostics panel (F). This panel contains several tabs to be inspected, which are shown in Figure 8a-8c. These show for each enabled group the number of open backdoor paths (VR12) and possible adjustment sets, the adjustment sets, and open backdoor paths, respectively (VR12)(VR14). Double-clicking one of the adjustment sets shown in Figure 8b will cause RoA to update the checkboxes shown in the panel accordingly(C). Double-clicking one of the (open) backdoor paths, shown in Figure 8c, will cause RoA to highlight the backdoor path in red in panel (E) (VR13).

While interacting with the graph it is helpful to use import or export using the top menu (VR7). To minimize clutter, one of the built-in automatic layout algorithms can be used Chimani et al. (2014): orthogonal, energy-based, and Sugiyama layouts, which

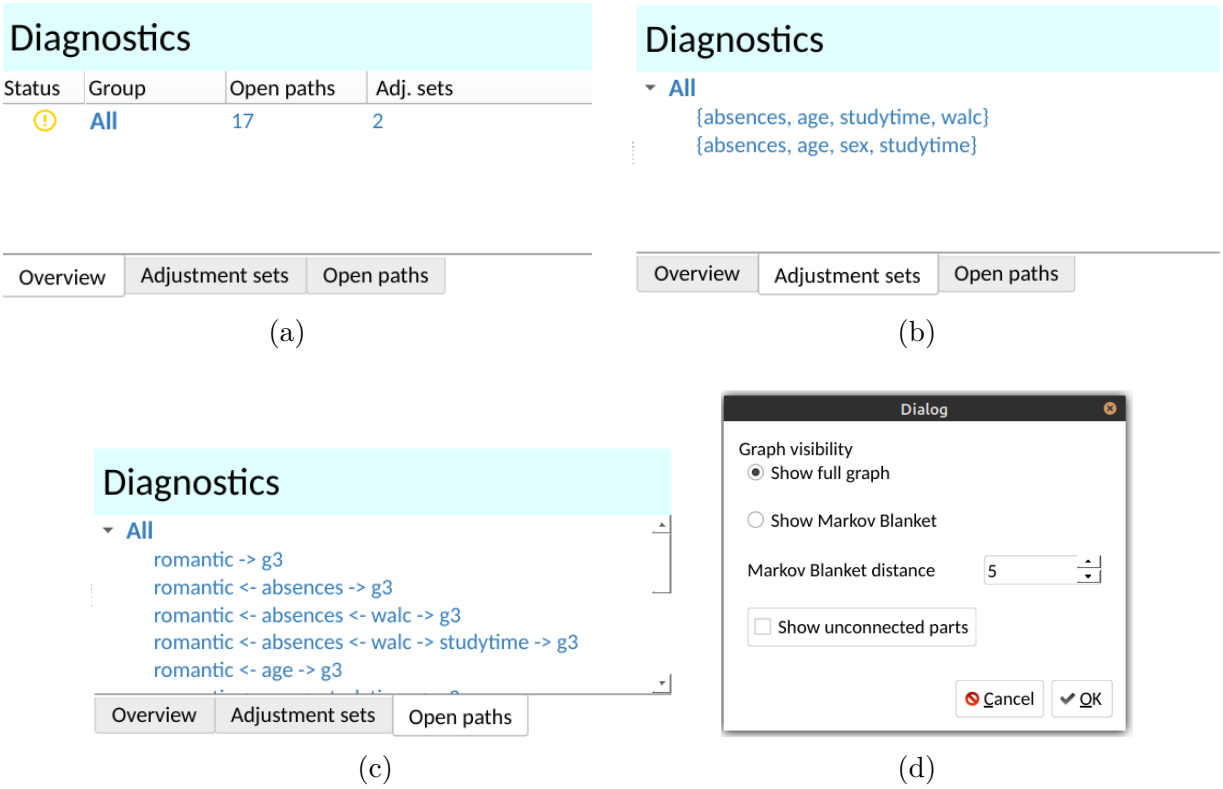


Figure 8: (a) Diagnostics panel showing overview of backdoor paths and adjustment sets. The yellow icons indicates that backdoor paths were found, but that adjustments are possible. A red icon would indicate backdoor paths with no possible adjustment sets. A green icon would indicate that no backdoor paths were found. (b) Diagnostics panel showing the possible adjustment sets. (c) Diagnostics panel showing the found backdoor paths. (d) Dialog for selecting viewing options for the causal graph.

are accessible via the "Layout" button on the top-right side of panel (E)(VR9). In addition, we have added two more options that are accessible through a dialog that pops up when clicking the "View options" button, which is shown in Figure 8d (VR10). The first option selects whether the full graph is shown or a reduced one based on the *Markov Blanket* that we will explain shortly. The second option is to select whether unconnected nodes, representing variables, are shown or hidden.

For a given causal relationship under study, the (minimal) Markov Blanket (MB) is a set of nodes of the graph that correspond to variables that are relevant for deconfounded effect estimation. In other words, these are the nodes to consider when searching for open backdoor paths. During expert discussions and while editing the causal graph, it may be convenient to consider nodes that are not the MB but are still close to in terms of graph distance. For this reason, the researcher can select a distance threshold to view nodes close the boundary of the MB as well, see Figure 9. Graph reduction based on the MB is used for our use case described in Section 3.

Thus, using the panels discussed so far, the researcher can interactively select a causal relationship of interest, manually balance variables, update the causal graph embedding the relationship, monitor open backdoor paths and counter the confounding effects allowed for by open backdoor-paths by balancing variables based on computed adjustment

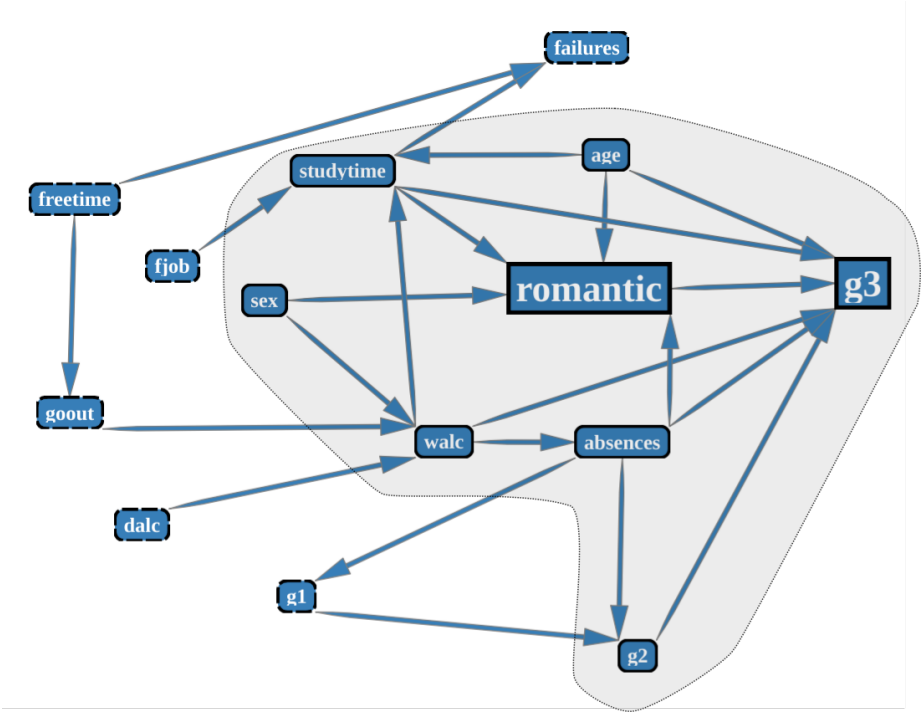


Figure 9: The Markov blanket for effect variable *g3*. All variables that are relevant for estimating a causal effect size on *g3* are inside the Markov blanket, indicated by the shaded bounded area. Variables outside this area have a certain distance to the Markov blanket in terms of edges (arrows) in between. For instance the variable *fjob*, which indicates the job of the father, has a distance of 1 to the Markov blanket, while the variable *freetime* has a distance of 2.

sets if these exist. During this process, the treatment effect estimates with associated distributions and Glass's delta values, are updated in real-time and shown in panel (C) (VR16)(VR17). If variables have been balanced, the effect estimated is the adjusted one. The visualizations used in these panels corresponds to the ones shown in Figure 5.

For our example, the ATE estimand option would measure how the phenomenon of having romantic relationships in general changes the average math grade on a population level. As shown in Figure 10, the effect is bigger than for the ATT (0.52 instead of 0.17, with Glass's delta 0.249 instead of 0.0047) and the effect distributions are different in a statistically significant manner, as indicated by the red background color behind the Glass's delta text. Also the SMD values are different in the balance panel.

Finally, we have implemented support for comparing subgroups during the analysis. Comparison of subgroups can be helpful during debates to eliminate the confounding effects of a variable (that cannot be adequately adjusted for), a scenario included in Figure 7. For instance, if the variable *sex* turns out to be problematic, subgroups can be defined based on it to eliminate its confounding effects. This yields two separate models for males and females.

Once defined, RoA associates each subgroup with two similar colors for visualizations and individual causal models, whereafter individual effect estimates are computed. The panels are then updated to contrast group sizes, variable distributions, SMD values,

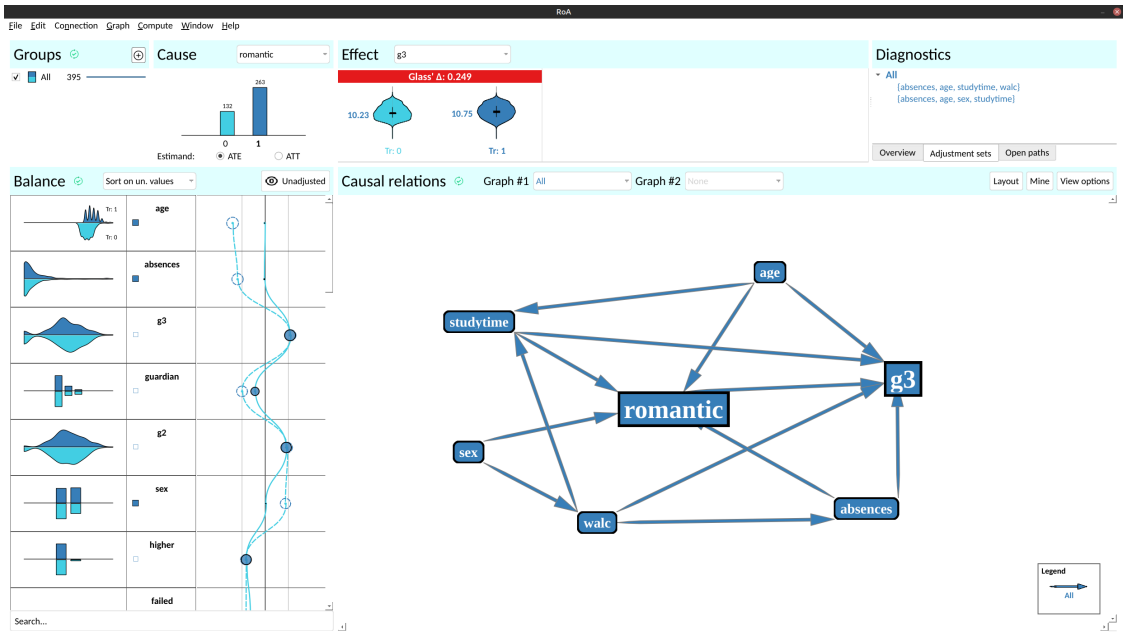


Figure 10: The effect of romantic relationships on the average math grade on a population level, estimated using the ATE.

effect sizes, and diagnostics, as shown in Figure 11. Furthermore, variables can be adjusted using checkboxes on a group basis in the "Balance"-panel on the left. The "Causal relations"-panel offers a pairwise comparison of graphs. Each subgroup has a causal graph associated that can be compared with the causal graph of another subgroup or with the outputted graph of the mining algorithm for the same group. These graphs to be compared can be selected using the "Graph #1" and "Graph #2" comboboxes. The first is used for layout algorithms and view options; the second is shown using a difference visualization (VR11).

To support the visual comparison of two causal diagrams, we have experimented with different edge encodings (see Figure 12a). In the end, we settled on the encodings shown in Figure 12b for comparing the causal diagram of a group with the automatically mined causal diagram. Similarly, for comparing two different subgroups, we settled on the encodings shown in Figure 12c. We focussed on showing differences with minimal clutter; therefore, when two causal diagrams both include an edge (arrow), it was colored to appear more in the background. When two causal diagrams disagree on the direction of an edge, the edge is drawn bigger to draw more attention to it.

Arrows are colored to highlight differences, see Figure 12b. Arrows colored with a primary group color are only part of the graph of that group and white arrows indicate that the two graphs both include an arrow. Again, when the two graph include arrows of opposite direction these are drawn close together and a little thicker to make them stand out more. In this way, the differences between the graphs are emphasized. This technique is also used to compare the graph of a group with an automatically mined graph (see Section 3). The arrows in the mined graph are then drawn in dark gray, see Figure 12c.

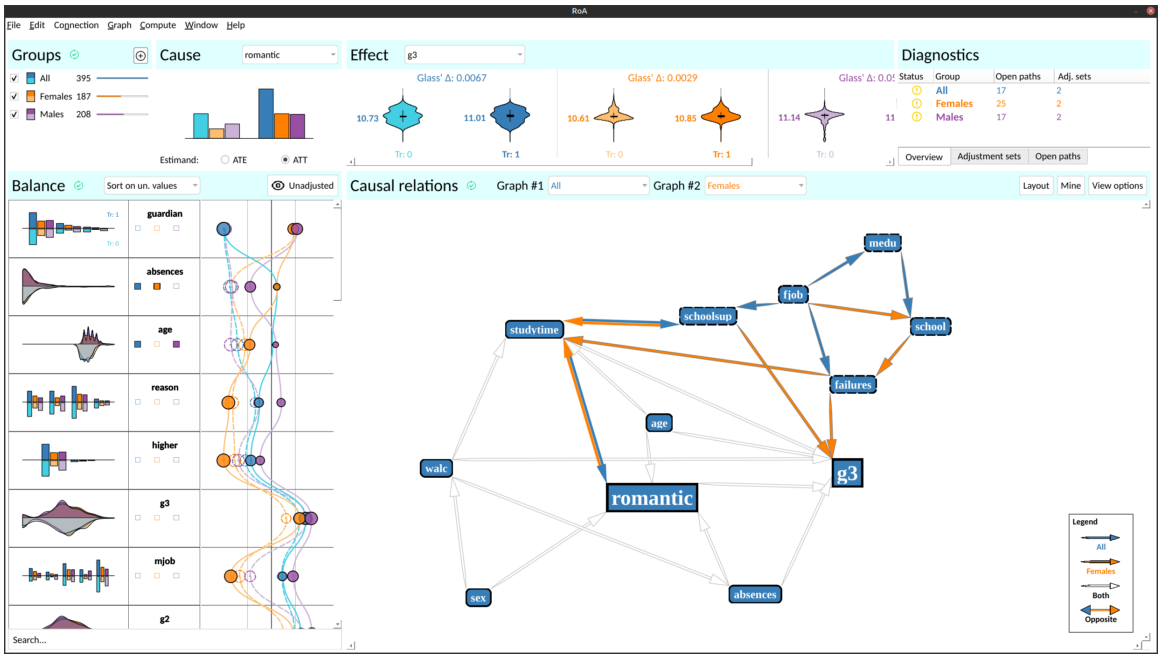


Figure 11: RoA supporting comparison of subgroups.

3. Predicting recovery time after epidural administration

In this section, we discuss our use case for a real-world experiment. The research team consisted of five members: one data scientist, three clinical researchers (usually working as anesthesiologists), and a statistician, with the first one interacting with the software during the discussions.

Data and research question

For our use case, we used a data set that was compiled by clinical researchers in the intensive care unit of our local hospital (with approval by the institutional review board). This data set contains over 21k records of patients that underwent surgery and includes 80 variables related to their condition and recovery process. Our research objective was to estimate the effect epidural administration (anesthesia using nerve blocking injections) on the recovery time in minutes ($epidural \rightarrow time\_min\_recovery$), motivated by improving post-operative planning. For this study endpoint, we used the ATT estimand.

Constructing the causal graph

A causal graph was constructed for a related project, so we could take that as a starting point. The full graph is shown in Figure 13a. After selecting the cause and effect variables and importing the expert graph, we employed our mining algorithm (see Appendix A) to automatically mine more edges (arrows). Using the causal relations panel we compared the expert graph with the mined graph, see Figure 13b. It was found that most mined edges were in less relevant parts of the graph. Typically, when explicit disagreement was found between the graphs, the experts could dismiss the mined edges based on clinical reasoning. Some small discussion followed on particular edges, but

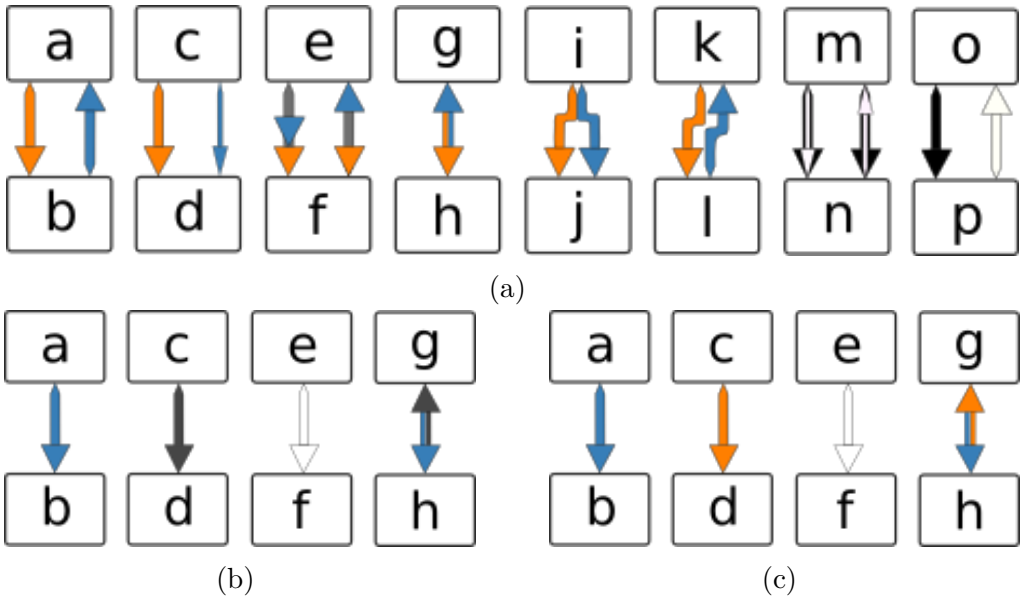


Figure 12: Directed edge (arrows) color encoding for graph comparison. (a) Experimental encodings. (b) Edges associated for groups obtain the group color, indicated by $a \rightarrow b$. Edges mined with a mining algorithm based on the data are colored gray, see $c \rightarrow d$. If the graph of a group and the mined graph agree on an edge this is indicated with a white color to make the edge less visually striking, see $e \rightarrow f$. If the group graph and mined graph disagree this is indicated by a bi-directional edge, like $g \rightarrow h$. These type of edges are also drawn a bit larger to make them stand out more. (c) The coloring is analogous to that of (b) but now for the causal graphs of two different subgroups, using their primary group colors.

after reducing the expert causal graph using the Markov Blanket, the experts agreed on the graph (see Figure 13c–13d).

Balance check

During discussions the SMD value of variables under consideration were inspected to gauge their potential impact using the balance panel, see in Figure 14 and Figure 15. RoA found 169 open backdoor paths and one adjustment set: {*age*, *bmi*, *spoed*(urgency), *surgery_group*}. The adjustment set was selected and used to update the effect estimates.

Effect estimation

Without adjustment, the average recovery time was close to 70 minutes for both distributions. After applying the adjustment set these changed to 77.45 and 76.77, for patients without and with epidural administration, respectively. The two distributions are statistically significantly different, but based on the difference of 0.68 minutes, we conclude given this dataset, that epidural administration does not have a clinically relevant effect on recovery time.

To show that in an OBS the ATE and ATT indeed lead to two different estimates we selected the ATE option in the treatment panel and recompute the estimates, see

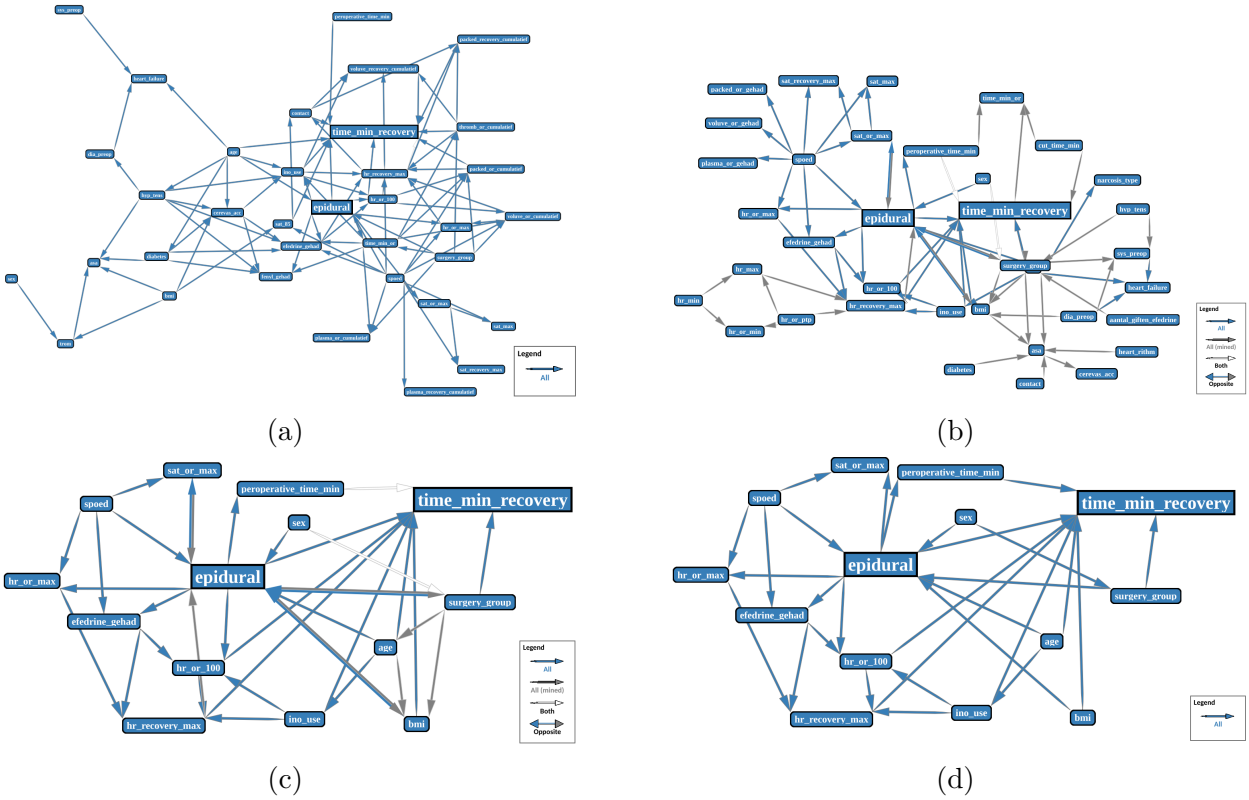


Figure 13: (a) The full expert causal graph. (b) The full expert graph compared with the mined causal graph after applying the energy-based graph layout algorithm. (c) The expert graph reduced using its Markov Blanket, while being compared with the mined graph. (d) The expert graph reduced using its Markov Blanket.

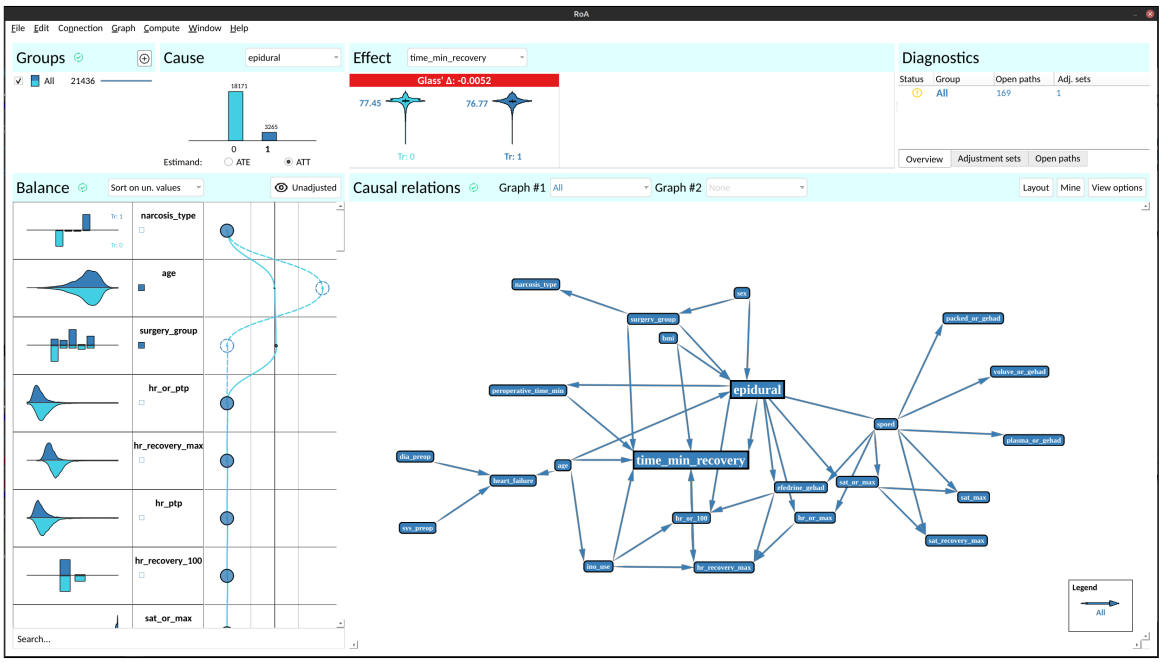
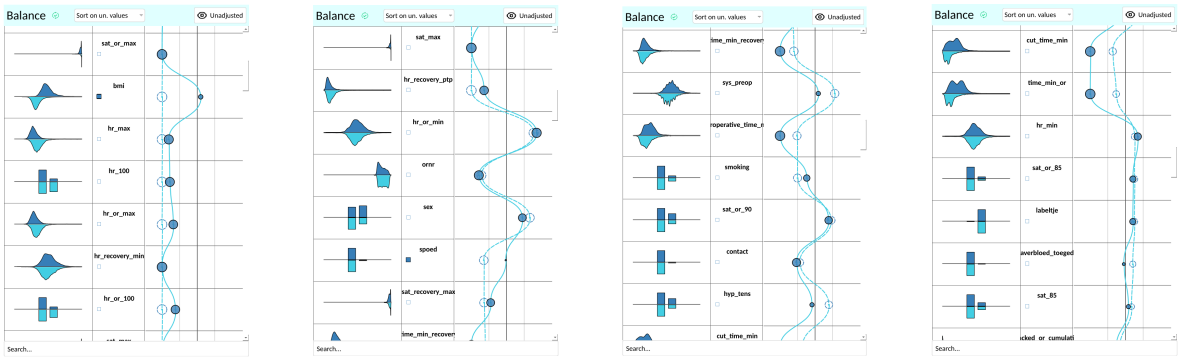


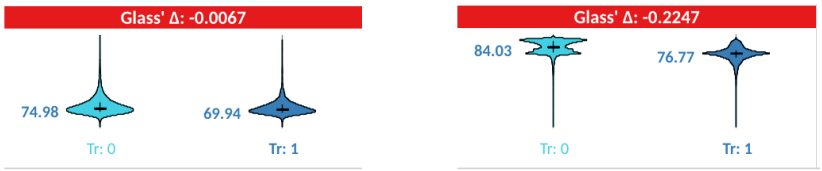
Figure 14: Using RoA for predicting recovery time after epidural administration.



(a) (b) (c) (d)

Figure 15: (a-d) Different parts of the balance panel after being scrolled. The plot confirms that the SMD values of the variables in the adjustment set are now indeed closer to 0.

Figure 16.



(a) (b)

Figure 16: (a) The ATE estimated without applying the adjustment set. (b) The ATE estimated after applying the adjustment set.

During the experiment, we collected user feedback which can be found in Appendix C. Furthermore, an additional use case can be found in the supplementary materials of this work.

4. Discussion

We have learned many lessons during the development of RoA. Overall, user feedback was positive. After a quick tutorial session, the tool was regarded as a welcome addition to the clinical research toolkit. A major reason is that clinical researchers value their time and like to adopt tools that efficiently and adequately guide them in their studies' exploratory phase. By integrating robust statistics in an exploratory setting, the researchers could immediately start exploring hypotheses in an intuitive way with a high level of confidence. Furthermore, by making the domain knowledge explicit in the form of a causal graph, differences in opinion are either resolved or lead to different hypotheses and implications that can be contrasted quickly. In this light, they underlined the importance of having simple intuitive UI's in which complexity is only added if necessary for the task. In the remainder, we discuss some issues and suggest improvements for future work.

Causal graphs

We have added automated support for mining causal graphs in RoA. Because the algorithms have a non-deterministic component, the mined graph differs after every run. Running the algorithm multiple times and using an aggregated result helps, but in our experience the value of mined graphs was limited, because clinical researchers tend to rely more on their own opinion. Furthermore, smarter layout algorithms could be adopted that are better suited for path analysis (Mennens et al. 2019). Based on this, different types of relevant substructures could be highlighted in conjunction with automated suggestions on how to explore different alternative CNs, depending on the context. These might be just as important as showing the backdoor paths.

In our system, edges are either present or not in the expert graph by design, which is required to compute the backdoor paths correctly (for mined graphs, we can already set a cutoff value for certainty). However, showing on-demand information about certainty on or next to the edge being considered may still be useful, along with expert annotations. Another extension worth considering is adding support for detecting plausible "front-door" paths, which can allow for treatment effect estimation in some cases when backdoor paths cannot be adjusted (Pearl et al. 2016c).

Variables

A challenge that kept emerging when developing causal networks was interpreting pre-recorded variables correctly. Different people tend to use different encodings. For example, one might split a variable into two variables to represent two different measurements over time. This can influence the definition of a causal network. Additionally, two phenomena that mutually influence each other, like drugs $\leftrightarrow$ heart rate, need to be rolled out to adhere to the assumption of causal networks being DAGs. Hence, proper data preparation is necessary.

Support for guided subgroups selection

It could be beneficial to integrate guided subgroup selection based on group similarity measures using a visual partition diagram (Gotz et al. 2017). We recognize that the number of groups that can be enabled and shown simultaneously without causing too much clutter or exhausting the available colors is limited to around three.

Future work

We envision direct extensions of the tool to include guided (sub)group selection mechanisms for the detection and utilization of front-door paths, smarter graph layout algorithms for pathfinding and substructure highlighting, encoding for edge uncertainty (in strength and direction) and user annotations, the support of different types of treatment and effect variables, and perhaps other propensity adjustment methods, like matching and stratification.

5. Conclusion

We have presented a visual analytics system for exploratory observational studies to

estimate the causal effect of a binary treatment variable. The system comprises several interactive plots that work cohesively together to integrate state-of-the-art statistical techniques to cover different aspects of causal inference, like differences in variable distributions between treatment groups, the structure of relevant causal relationships, and proper adjustment sets of variables used to obtain deconfounded treatment effect estimations when possible. Furthermore, automatic mining algorithms were integrated into the system to aid clinical researchers in finding the structure of the relevant causal relationships.

The system's functionality was demonstrated through a real-world use case involving patients that underwent surgery, conducted in close collaboration with clinical researchers. In particular, the interactive capability of the system enabled clinical researchers to obtain statistically sound causal effects immediately during discussions about the causal relations relevant to the case. Because the causal relations were shown explicitly, the researchers could focus on efficiently exploring hypotheses together, which they regarded as a welcome aid to their clinical practice. During the sessions, the researchers tended to rely solely on their domain knowledge than the (non-deterministic) results of the mining algorithm.

Computational Details

The main application for RoA was programmed in C++ 9.4 using the Qt 6.2 framework (https://www.qt.io/). The statistical computations were programmed in R 4.1.0 using the following packages (https://CRAN.R-project.org/):

- plumber 1.1.0 (web API for R);
- jsonlite 1.7.2 (parsing JSON);
- RPostgreSQL 0.6-2 (database connection);
- cobalt 4.3.1 (covariate balance evaluation);
- survey 4.0 (effect estimation);
- dagitty 0.3-1 (computing of adjustment sets);
- MXM 1.5.1 (computing of causal models).

The code can be found on GitHub at https://github.com/RodofAsclepius/RoA.

Computation time

For our use case (and experiments), we selected 21k subjects and 80 variables. This allowed for interactive use of the tool due to the computational load on a contemporary desktop system running a local *R* server. Although we already selected algorithms on the more efficient side of the spectrum, it became clear that graph mining should be supported with dedicated (cloud) servers for use cases involving bigger datasets. The time required by the mining algorithm is also highly sensitive to its parameters involving edge certainty.

Acknowledgments

This work is a result of the Eindhoven MedTech Innovation Center (e/MTIC 2021).

References

Abadie, A. and Cattaneo, M. D. (2018). Econometric methods for program evaluation. *Annu. Rev. Econ*, 10:465–503, DOI: 10.1146/annurev-economics, https://doi.org/10.1146/annurev-economics-.

Agresti, A. (2012). *Categorical Data Analysis, 3rd Edition*. Wiley, Hoboken, NJ, USA, ISBN: 978-0-470-46363-5, https://www.wiley.com/en-us/Categorical+Data+Analysis%2C+3rd+Edition-p-9780470463635.

Austin, P. C. (2009). Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in Medicine*, 28:3083–3107, ISSN: 02776715, DOI: 10.1002/sim.3697.

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. *Multivariate Behavioral Research*, 46(3):399–424, ISSN: 00273171, DOI: 10.1080/00273171.2011.568786.

Bae, J., Helldin, T., and Riveiro, M. (2017a). Understanding Indirect Causal Relationships in Node-Link Graphs. *Computer Graphics Forum*, 36(3):411–421, ISSN: 14678659, DOI: 10.1111/cgf.13198.

Bae, J., Ventocilla, E., Riveiro, M., Helldin, T., and Falkman, G. (2017b). Evaluating multi-attributes on cause and effect relationship visualization. *VISIGRAPP 2017 - Proceedings of the 12th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications*, 3(Visigrapp):64–74, ISBN: 9789897582288, DOI: 10.5220/0006102300640074.

Beal, S. J. and Kupzyk, K. A. (2014). An Introduction to Propensity Scores: What, When, and How. *Journal of Early Adolescence*, 34(1):66–92, ISBN: 0272431613, ISSN: 02724316, DOI: 10.1177/0272431613503215.

Chimani, M., Gutwenger, C., Jünger, M., Klau, G., Klein, K., and Mutzel, P. (2014). The open graph drawing framework (ogdf). In Tamassia, R., editor, *Handbook of Graph Drawing and Visualization*, chapter 17. CRC Press.

Collins, R., Bowman, L., Landray, M., and Peto, R. (2020). The magic of randomization versus the myth of real-world evidence. *New England Journal of Medicine*, 382(7):674–678, ISSN: 15334406, DOI: 10.1056/NEJMsb1901642.

Cortez, P. and Silva, A. (2008). Using Data Mining To Predict Secondary School Student Performance. *EUROSIS*.

Dang, T. N., Murray, P., Aurisano, J., and Forbes, A. G. (2015). ReactionFlow: An interactive visualization tool for causality analysis in biological pathways. *BMC Proceedings*, 9(Suppl 6):S6, ISSN: 17536561, DOI: 10.1186/1753-6561-9-S6-S6, http://www.biomedcentral.com/1753-6561/9/S6/S6.

Elmqvist, N. and Tsigas, P. (2003). Causality visualization using animated Growing Polygons. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, 2003:189–196, ISBN: 0780381548, ISSN: 1522404X, DOI: 10.1109/INFVIS.2003.1249025.

e/MTIC (2021). *Eindhoven Medtech innovation center*, https://www.emtic.nl/.

Fonarow, G. C. (2016). Randomization - There is no substitute. *JAMA Cardiology*, 1(6):633–635, ISSN: 23806591, DOI: 10.1001/jamacardio.2016.1792.

Fox, J. (2008). *Applied Regression Analysis and Generalized Linear Models*. SAGE Publications, ISBN: 9780761930426, http://books.google.ch/books?id=GKkn3LSSHFsC.

Ghoniem, M., Fekete, J. D., and Castagliola, P. (2004). A comparison of the readability of graphs using node-link and matrix-based representations. *Proceedings - IEEE Symposium on Information Visualization, INFO VIS*, pages 17–24, ISBN: 0780387793, ISSN: 1522404X, DOI: 10.1109/INFVIS.2004.1.

Gotz, D., Sun, S., Cao, N., Kundu, R., and Meyer, A. M. (2017). Adaptive contextualization methods: For combating selection bias during high-dimensional visualization. *ACM Transactions on Interactive Intelligent Systems*, 7(4):1–23, ISSN: 21606463, DOI: 10.1145/3009973, http://dl.acm.org/citation.cfm?doid=3166060.3009973.

Greifer, N. (2022). *cobalt: Covariate Balance Tables and Plots*, https://CRAN.R-project.org/package=cobalt. R package version 4.4.1.

Guo, H., Huang, J., and Laidlaw, D. H. (2015). Representing Uncertainty in Graph Edges: An Evaluation of Paired Visual Variables. *IEEE Transactions on Visualization and Computer Graphics*, 21(10):1173–1186, ISSN: 10772626, DOI: 10.1109/TVCG.2015.2424872.

Hill, A. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*.

Ho, D. E., Imai, K., King, G., and Stuart, E. A. (2007). Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. *Political Analysis*, 15(3):199–236, ISSN: 10471987, DOI: 10.1093/pan/mpl013.

Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association*, 81(396):945–960, ISSN: 1537274X, DOI: 10.1080/01621459.1986.10478354.

Huang, W., Li, S., and Peng, L. (2022). Extreme continuous treatment effects: Measures, estimation and inference. http://arxiv.org/abs/2209.00246.

Imbens, G. W. and Wooldridge, J. M. (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1):5–86, https://ideas.repec.org/a/aea/jeclit/v47y2009i1p5-86.html.

Jin, Z., Guo, S., Chen, N., Weiskopf, D., Gotz, D., and Cao, N. (2021). Visual Causality Analysis of Event Sequence Data. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1343–1352, ISSN: 19410506, DOI: 10.1109/TVCG.2020.3030465.

Kadaba, N. R., Irani, P. P., and Leboe, J. (2007). Visualizing causal semantics using animations. *IEEE Transactions on Visualization and Computer Graphics*, 13(6):1254–1261, ISSN: 10772626, DOI: 10.1109/TVCG.2007.70528.

Kalisch, M., Mächler, M., Colombo, D., Maathuis, M. H., and Bühlmann, P. (2012). Causal inference using graphical models with the R package pcalg. *Journal of Statistical Software*, 47(11), ISSN: 15487660, DOI: 10.18637/jss.v047.i11.

Kang, J. D. Y. and Schafer, J. L. (2007). Demystifying Double Robustness: A Comparison of Alternative Strategies for Estimating a Population Mean from Incomplete Data. *Statistical Science*, 22(4):523 – 539, DOI: 10.1214/07-STS227, https://doi.org/10.1214/07-STS227.

Lagani, V., Athineou, G., Farcomeni, A., Tsagris, M., and Tsamardinos, I. (2017). Feature selection with the R package MXM: Discovering statistically equivalent feature subsets. *Journal of Statistical Software*, 80(7), DOI: 10.18637/jss.v080.i07.

Ledesma, R. D., Macbeth, G., and de Kohan, N. C. (2009). Computing effect size measures with vista-the visual statistics system. *Tutorials in Quantitative Methods for Psychology*, 5:25–34, DOI: 10.20982/tqmp.05.1.p025.

Leite, W. (2016a). *Practical Propensity Score Methods Using R*, pages 8–9. ISBN: 9781452288888, DOI: 10.4135/9781071802854.

Leite, W. (2016b). *Practical Propensity Score Methods Using R*, pages 1–68. ISBN: 9781452288888, DOI: 10.4135/9781071802854.

Leite, W. (2016c). *Practical Propensity Score Methods Using R*, pages 62–63. ISBN: 9781452288888, DOI: 10.4135/9781071802854.

Leite, W. (2016d). *Practical Propensity Score Methods Using R*, pages 62–63. ISBN: 9781452288888, DOI: 10.4135/9781071802854.

Lemon, J. (2006). Plotrix: a package in the red light district of r. *R-News*, 6(4):8–12.

Loh, W. W. and Vansteelandt, S. (2020). Confounder selection strategies targeting stable treatment effect estimators. DOI: 10.1002/sim.8792, http://arxiv.org/abs/2001.08971http://dx.doi.org/10.1002/sim.8792.

Lumley, T. (2020). survey: analysis of complex survey samples. R package version 4.0.

Mennens, R. J., Scheepens, R., and Westenberg, M. A. (2019). A stable graph layout algorithm for processes. *Computer Graphics Forum*, 38(3):725–737, ISSN: 14678659, DOI: 10.1111/cgf.13723.

Nobre, C., Meyer, M., Streit, M., and Lex, A. (2019). The state of the art in visualizing multivariate networks. *Computer Graphics Forum*, 38(3):807–832, ISSN: 14678659, DOI: 10.1111/cgf.13728.

Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., and Gama, J. (2022). Methods and tools for causal discovery and causal inference. ISSN: 19424795, DOI: 10.1002/widm.1449.

Pearl, J., Glymour, M., and Jewell, N. (2016a). *Causal Inference in Statistics: A Primer*, pages 53–70. Wiley, ISBN: 9781119186847, www.wiley.com/go/Pearl/Causality.

Pearl, J., Glymour, M., and Jewell, N. (2016b). *Causal Inference in Statistics: A Primer*, pages 53–88. Wiley, ISBN: 9781119186847, www.wiley.com/go/Pearl/Causality.

Pearl, J., Glymour, M., and Jewell, N. (2016c). *Causal Inference in Statistics: A Primer*, pages 66–70. Wiley, ISBN: 9781119186847, www.wiley.com/go/Pearl/Causality.

Pearl, J., Glymour, M., and Jewell, N. (2016d). *Causal Inference in Statistics: A Primer*, pages 35–51. Wiley, ISBN: 9781119186847, www.wiley.com/go/Pearl/Causality.

Pearl, J. and Mackenzie, D. (2018). *The book of why: The new science of cause and effect*, page 13. Basic Books, ISBN: 9780465097609, http://bayes.cs.ucla.edu/WHY.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, https://www.R-project.org/.

Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, ISSN: 00063444, DOI: 10.1093/biomet/70.1.41.

Rubin D. B (1974). Estimating causal effects of treatment in randomized and nonrandomized studies. *Journal of Educational Psychology*, 66(5):688–701, http://www.fsb.muohio.edu/lij14/420_paper_Rubin74.pdf.

Schafer, J. and Kang, J. (2009). Average causal effects from nonrandomized studies: A practical guide and simulated example. *Psychological methods*, 13:279–313, DOI: 10.1037/a0014268.

Schafer, J. L. and Kang, J. (2008). Average Causal Effects From Nonrandomized Studies: A Practical Guide and Simulated Example. *Psychological Methods*, 13(4):279–313, ISSN: 1082989X, DOI: 10.1037/a0014268.

Shadish, W. R. (2010). Campbell and Rubin: A Primer and Comparison of Their Approaches to Causal Inference in Field Settings. *Psychological Methods*, 15(1):3–17, ISSN: 1082989X, DOI: 10.1037/a0015916.

Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. *Statistical Science*, 25(1):1–21, ISSN: 08834237, DOI: 10.1214/09-STS313.

Stuart, E. A. and Rubin, D. B. (2007). Best Practices in Quasi-Experimental Designs: Matching Methods for Causal Inference. *Best Practices in Quantitative Methods*, pages 155–176, DOI: 10.4135/9781412995627.d14.

Talbot, D., Diop, A., Lavigne-Robichaud, M., and Brisson, C. (2021). The change in estimate method for selecting confounders: A simulation study. *Statistical Methods in Medical Research*, 30:2032–2044, ISSN: 14770334, DOI: 10.1177/09622802211034219.

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T. (2016). Robust causal inference using directed acyclic graphs: the r package 'dagitty'. *International Journal of Epidemiology*, 45(6):1887–1894, DOI: 10.1093/ije/dyw341.

Tikka, S. and Karvanen, J. (2017). Identifying causal effects with the R package causaleffect. *Journal of Statistical Software*, 76(12), ISSN: 15487660, DOI: 10.18637/jss.v076.i12.

Tsagris, M., Borboudakis, G., Lagani, V., and Tsamardinos, I. (2018). Constraint-based causal discovery with mixed data. *International Journal of Data Science and Analytics*, 6(1):19–30, ISSN: 2364-415X, DOI: 10.1007/s41060-018-0097-y, https://doi.org/10.1007/s41060-018-0097-y.

VanderWeele, T. J. (2019). Principles of confounder selection. *European Journal of Epidemiology*, 34(3):211–219, ISBN: 0123456789, ISSN: 15737284, DOI: 10.1007/s10654-019-00494-6, https://doi.org/10.1007/s10654-019-00494-6.

Vogogias, A., Kennedy, J., Archambault, D., Bach, B., Anne Smith, V., and Currant, H. (2018). Bayespiles: Visualisation support for Bayesian network structure learning. *ACM Transactions on Intelligent Systems and Technology*, 10(1):1–23, ISSN: 21576912, DOI: 10.1145/3230623.

Wang, J. and Mueller, K. (2016). The Visual Causality Analyst: An Interactive Interface for Causal Reasoning. *IEEE Transactions on Visualization and Computer Graphics*, 22(1):230–239, ISSN: 10772626, DOI: 10.1109/TVCG.2015.2467931.

Wang, J. and Mueller, K. (2017). Visual Causality Analysis Made Practical. *IEEE Conference on Visual Analytics Science and Technology (VAST)*, pages 151–161.

Westreich, D., Lessler, J., and Funk, M. J. (2010). Propensity score estimation: neural networks, support vector machines, decision trees (CART), and meta-classifiers as alternatives to logistic regression. *Journal of Clinical Epidemiology*, 63(8):826–833, ISSN: 08954356, DOI: 10.1016/j.jclinepi.2009.11.020, http://dx.doi.org/10.1016/j.jclinepi.2009.11.020.

Witte, J. and Didelez, V. (2019). Covariate selection strategies for causal inference: Classification and comparison. *Biometrical journal. Biometrische Zeitschrift*, 61(5):1270—1289, ISSN: 0323-3847, DOI: 10.1002/bimj.201700294, https://doi.org/10.1002/bimj.201700294.

Xie, X., Du, F., and Wu, Y. (2021). A Visual Analytics Approach for Exploratory Causal Analysis: Exploration, Validation, and Applications. *IEEE Transactions on Visualization and Computer Graphics*, 27(2):1448–1458, ISSN: 19410506, DOI: 10.1109/TVCG.2020.3028957.

Xie, X., He, M., and Wu, Y. (2020). CausalFlow: Visual Analytics of Causality in Event Sequences. *ArXiv*, abs/2008.1, http://arxiv.org/abs/2008.11899.

Yen, C. H. E., Parameswaran, A., and Fu, W. T. (2019). An exploratory user study of visual causality analysis. *Computer Graphics Forum*, 38(3):173–184, ISSN: 14678659, DOI: 10.1111/cgf.13680.

A. Causal effect estimation in observational studies

Observational studies are conducted to estimate the effect causal size of a phenomenon Z on another phenomenon Y, denoted as $Z \to Y$. The cause variable Z is often referred to as the *treatment* variable, while the influenced variable Y is referred to as the *effect* variable. The generic analysis workflow for conducting such a study, which is used for our work, is shown in Figure 17.

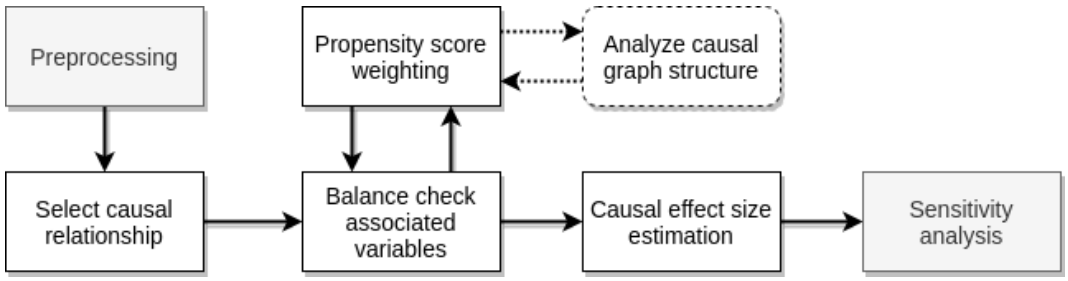


Figure 17: Workflow for estimating causal effects based on Leite (2016b) with additions for clarification and integrative purposes. The shaded rectangles indicate steps that are considered outside the main focus of this work.

The causal model for effect estimation

Several types of treatment effects can be estimated. To express these treatment effects, we adopt definitions formulated in terms of potential outcomes from Rubin's Causal Model (Rubin D. B 1974; Holland 1986; Shadish 2010). With this, we need to distinguish between observable and non-observable hypothetical outcomes. For a given subject i, the treatment indicator Z_i equals 1, if i is treated, and 0 otherwise. Correspondingly, for the (continuous) effect variable Y_i we denote the potential outcomes Y_i^z as follows.

$$Y_i^z = \begin{cases} Y_i^0, & \text{potential outcome without treatment} \\ Y_i^1, & \text{potential outcome with treatment.} \end{cases}$$

Furthermore, we define realized outcome $Y_i = Z_i \cdot Y_i^1 + (1 - Z_i) \cdot Y_i^0$ with observable outcomes (left) and non-observable potential outcomes, or *counterfactuals* (right)

$$Y_i = \begin{cases} Y_i^1 \mid Z_i = 1 \\ Y_i^0 \mid Z_i = 0 \end{cases} \quad Y_i = \begin{cases} Y_i^1 \mid Z_i = 0 \\ Y_i^0 \mid Z_i = 1. \end{cases}$$

Let us consider an example of counterfactual reasoning. Suppose you were concerned with driving home to be on time for dinner and had to pick road A or B. You picked A and arrived an hour late due to traffic congestion. Now you tell yourself that if you had picked road B instead, the driving time would have been shorter. Hence, now you are (re)assessing the driving time on road B while you have measured the driving time on road A.

Note that Y_i^1 and Y_i^0 denote the outcome of a treated and untreated subject i, which are invariant to assigned treatment condition Z_i. Given a subject i, the treatment effect is $\tau_i = Y_i^1 - Y_i^0$, but in reality one of course cannot observe Y_i^1 and Y_i^0 at the same

time. Therefore, we estimate the effect size using an estimand based on the expected values of the treated and untreated groups. Different types of estimands exist (Abadie and Cattaneo 2018; Nogueira et al. 2022; Huang et al. 2022), but the most common ones are listed in table 3. For a list of software packages to compute the estimands see Table 6.

Table 3: List of common estimands.

Name	**Target population**
The Average Treatment Effect (ATE)	Entire population
The Average Treatment Effect on the Treated (ATT)	Treated population
The Average Treatment Effect on the Control (ATC)	Control population
The Individual Treatment Effect (ITE)	Individual subject
Conditional Average Treatment Effect (CATE)	A subpopulation
Local Average Treatment Effect (LATE)	Compliers

We have picked on the ATE and the ATT because these effect measures are easy to interpret for a broad audience. Moreover, the ATT is often used in the medical field, so we chose to use it in our medical use case. These two effect measures are defined, respectively, as:

$$ATE = E(Y_i^1) - E(Y_i^0), \text{ and} \tag{1}$$
$$ATT = E(Y_i^1 \mid Z_i = 1) - E(Y_i^0 \mid Z_i = 1). \tag{2}$$

When conducting a Randomized Controlled Trials (RCT), the ATE, ATT, and ATC are all equal, but this is not the case for an OBS, where an analyst commonly picks the ATE or ATT depending on the research question. Because we are dealing with aggregates, we can estimate all terms of the equations by just including subjects that happen to be observed for the involved outcomes. Lastly, in order to obtain unbiased treatment effects, two assumptions are to be met:

1. The stable subject value assumption (SUTVA), meaning that potential outcomes for subject j are independent of subject i.

2. The strong ignorability assumption of treatment assignment (SITA), meaning that the treatment assignment is independent of potential outcomes distributions for a given set of observed covariates X or $(Y^0, Y^1) \perp Z \mid X$.

For our purposes, the main focus is on establishing assumption 2, while we assume assumption 1 to hold, which requires that for the probability of treatment assignment, for every value of the covariates, it holds that: $0 < p(Z_i = 1 \mid X) < 1$. The second assumption is required for unconfounded causal effect size estimation. Ideally, the distributions of all other covariates X, besides the treatment and effect variable, are balanced (informally: exhibit a high degree of equality or similarity) over the treatment groups. While randomization takes care of this in an RCT, we may need to manually adjust distributions for observational studies because, with no a priori control over the assignment of subjects to the treatment groups, randomization is impossible. For this, we can utilize techniques such as weighting, matching, and stratification to obtain

strong evidence that the second assumption is being met. For our work, we have adopted the use of propensity score weighting.

Propensity score weighting

The imbalance in the distributions of covariates X across treatment subgroups can be summarized with scalar valued scores per subject. An increasingly popular choice for such a score is the propensity score that is defined as the conditional probability of treatment assignment (Rosenbaum and Rubin 1983):

$$e_i(X) = P(Z_i = 1 \mid X). \tag{3}$$

The propensity score calculation can be done with a multitude of methods (Westreich et al. 2010), but in the case of a dichotomous treatment, the standard method is to use a logistic regression model (Agresti 2012; Fox 2008), which practice we follow. The logistic regression model for estimating propensity scores is defined as

$$\text{logit}(Z_i = 1 \mid X) = \beta_0 + \beta_1 X_{i1} + \cdots + \beta_k X_{ik}, \tag{4}$$

where the logit equals the log odds of the probability of getting the treatement:

$$\text{logit}(Z_i = 1 \mid X) = \log\left(\frac{P(Z_i = 1)}{1 - P(Z_i = 1)}\right). \tag{5}$$

The probabilities of treatment assignments $e_i(X)$, given the covariates X, are then estimated using the logits:

$$e_i(X) = P(Z_i = 1 \mid X) = \frac{\exp(\text{logit}(Z_i = 1 \mid X))}{1 + \exp(\text{logit}(Z_i = 1 \mid X))}. \tag{6}$$

By integrating the propensity scores in a list of weights W (one per subject), it is possible to obtain weighted distributions via weighted means or weighted proportions that are likely to be more statistically similar in the treated and untreated groups of subjects. In this way, the propensity score is used to establish the SITA assumption. Specifically we have $(Y^0, Y^1) \perp Z \mid X \implies (Y^0, Y^1) \perp Z \mid e(X) \wedge Z \perp X \mid e(X)$. Hence, also the treatment assignment becomes independent of X.

The actual method to be used for propensity score weighting depends on the type of treatment effect one wants to estimate. To estimate the weights W for the ATE (Equation 1) and the ATT (Equation 2) one needs to compute the weights with:

$$\text{ATE case:} \quad w_i = \frac{Z_i}{e_i(X)} + \frac{1 - Z_i}{1 - e_i(X)}; \text{ and} \tag{7}$$

$$\text{ATT case:} \quad w_i = Z_i + (1 - Z)\frac{e_i(X)}{1 - e_i(X)}. \tag{8}$$

The remaining question is: which variables to include in X to establish the premise? The misspecification of the propensity score model by including an unsuitable set of covariates can lead to substantial bias in the estimation of the causal effects. Current (clinical) guidelines often suggest adding a variable to X based on domain knowledge

and whether it is associated with Z and/or Y, which decreases bias and variance of the treatment estimations (Beal and Kupzyk 2014; VanderWeele 2019; Witte and Didelez 2019; Loh and Vansteelandt 2020; Talbot et al. 2021). Unfortunately, however, these guidelines are insufficient to ensure a valid X to make the premise hold, and analysts have to be pragmatic in checking the achieved balance and advocate for the plausibility of chosen X. However, the relatively recent work by Pearl on do-calculus does offer a theoretical basis for selecting the correct variables, which is covered below.

During the weighting process, it is important to assess the imbalance in distributions for each covariate in X by contrasting the treatment groups. For this, we use the Standardized Mean Difference (SMD), which is an established indicator for balance assessment. It is defined for continuous and dichotomous variables as (Austin 2009):

$$\text{Continuous case:} \quad d = \frac{(\bar{x}^1 + \bar{x}^0)}{\sqrt{\frac{(s^1)^2 + (s^0)^2}{2}}}; \text{ and} \tag{9}$$

$$\text{Dichotomous case:} \quad d = \frac{(\hat{p}^1 - \hat{p}^0)}{\sqrt{\frac{\hat{p}^1(1-\hat{p}^1) + \hat{p}^0(1-\hat{p}^0)}{2}}}, \tag{10}$$

where $\bar{x}^1$ and $\bar{x}^0$ denote the sample means for the covariate for the treated and untreated group, respectively, while s^1 and s^0 denote the sample variance of the continuous covariate for respectively the treated and treated group. Furthermore, $\hat{p}^1$ and $\hat{p}^0$ denote the prevalence (or mean) of the dichotomous covariate for the treated and untreated group, respectively. The SMD should be as close to zero as possible for optimal balance. In practice, a cut-off value must be chosen to classify covariates as balanced or unbalanced. A typically recommended cut-off value is 0.1 (Austin 2011) or 0.25 (Stuart and Rubin 2007; Stuart 2010). If, after weighting, some (relevant) covariates remain unbalanced, one can try to include higher-order / interaction terms in the propensity model or consider stratification on unbalanced covariates (subgroup analysis).

Causal do-calculus and graph mining

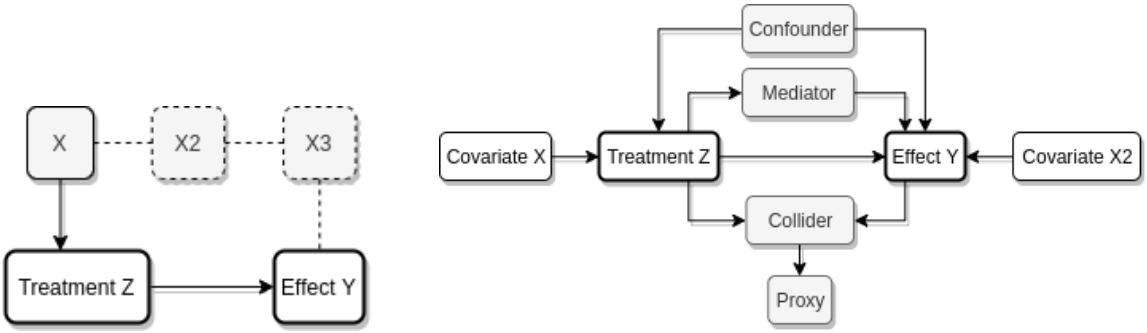
Causal do-calculus relies on domain knowledge, which can be regarded as a list of assumptions, but the representation thereof (e.g., logical statements, structural equations, diagrams) can make a profound difference. Nevertheless, causal diagrams are a sound option for nearly all applications because of their transparency and explicitness in answering questions, as stated by Pearl and Mackenzie (2018). Our adopted representation of a causal diagram is a Directed Acyclic Graph (DAG), although more complex variations exist (Pearl et al. 2016d; Kalisch et al. 2012). Furthermore, there are three steps or subareas in causal inference (Tikka and Karvanen 2017):

1. the discovery of the causal model (from data);
2. the identification of causal effects using a known model;
3. the actual estimation of an identified causal effect from data.

Only "identifiable" effects can be adequately estimated, for which it holds that conditional upon the correctness of the DAGs structure: the absence of confounding is guaranteed, the presence of confounding can be adjusted for, or the presence of confounding cannot be adjusted for. This is the case when specific relevant substructures, called *backdoor paths*, are either absent in the DAG or can be adjusted for. An example of a simple backdoor path is illustrated in Figure 18a.

Given a DAG containing the causal treatment-effect relationship of interest $Z \to Y$, a backdoor path is a sequence of connected nodes that connects Z and Y, such that Z is connected to the sequence via an incoming edge. One can consider these backdoor paths as a more complex variant, or generalization, of a confounding variable that would express itself as a node that is directly connected with directional edges pointing towards Z and Y. In the case of an RCT, the randomization process neutralizes possible confounding effects of backdoor paths by "cutting" all incoming arrows of Z. In the case of an OBS, we need to identify these backdoor paths using do-calculus and adjust for them if possible to obtain deconfounded treatment effect estimates Pearl et al. (2016a).

The sequence of connected nodes in a backdoor path does not have to be connected via consecutive edges, all pointing in the same direction along the path. Only the nodes connected to the treatment and effects variables need to point in the direction of the treatment and effect nodes, making the situation more complex than having only a straightforward path connected. Causal do-calculus is designed to work with different substructures like the four primary substructures compactly illustrated in Figure 18b: the confounder, mediator, collider, and the proxy variables. These are used for computing proper adjustment sets when backdoor paths are present.



(a) (b)

Figure 18: The relevant substructures in causal diagrams: (a) Illustration of a backdoor path. (b) The confounder, mediator, and collider are all associated with variables Z and Y, but the underlying causal directions differ. The proxy is relevant because adjusting for it means implicitly adjusting for a collider, which in turn can cause an open backdoor path to form.

Depending on the overall structure, it can be the case that: there are no open backdoor paths (which allows for immediate effect estimation), there are open backdoor paths, but these can be compensated for using a proper adjustment set of covariates, or there are unadjustable open backdoor paths (which implies that no unbiased effect estimation is possible). Naturally, the result is always conditional on the correctness of a given causal network.

Generally, a causal network is specified using domain knowledge, by guessing and checking, or by utilizing mining algorithms (see Table 5 in Appendix B). Earlier tests had been developed for either categorical or numerical variables only, but the method published by Tsagris et al. (2018) can readily handle varied situations. For RoA, we have deployed this method and injected it into the *PC* mining algorithm (Kalisch et al. 2012).

Causal effect estimation

For estimating a causal treatment effect using propensity score weighting, we use (Schafer and Kang 2008):

$$\Delta = \frac{\sum_{t=1}^{T} w_t y_t}{\sum_{t=1}^{T} w_t} - \frac{\sum_{u=1}^{U} w_u y_u}{\sum_{u=1}^{U} w_u}, \tag{11}$$

where w_t, y_t and T are the weights, *observed* outcomes and size, respectively, for the treated group, w_u, y_u and U are the weights, *observed* outcomes and size, respectively, for the untreated group. Please note that an individual subject is either part of the treated or untreated group. Therefore the effect is estimated with the difference in the weighted average of the outcomes in these groups. Depending on whether the ATE or ATT is required, the weights are computed using Equation 7 or Equation 8.

With balancing being performed in the study's design phase, one can pick any from a diverse set of techniques for treatment estimation Ho et al. (2007). Since we designed our tool for exploratory purposes, we have adopted the doubly robust method for treatment effect estimation (Kang and Schafer 2007; Schafer and Kang 2009), based on propensity scores, to minimize bias. A key characteristic of the doubly robust method is that if either the propensity model (treatment assignment mechanism) or the outcome (response) model is correct (concerning the real world), the treatment effect estimates are unbiased Imbens and Wooldridge (2009). Using this method the values y_t and y_u in Equation 11 are substituted with *predicted* values $\hat{y}_t$ and $\hat{y}_u$. To obtain these predicted values, we first fit two separate outcome models (Leite 2016d):

$$M_t = \beta_t + \beta_{1t} P_t + \beta_{2t} P_t^2 + \beta_{3t} P_t^3 + \epsilon_t; \text{ and} \tag{12}$$
$$M_u = \beta_u + \beta_{1u} P_u + \beta_{2u} P_u^2 + \beta_{3u} P_u^3 + \epsilon_u, \tag{13}$$

where M_t models the outcome for the treated subgroup, based on a linear, quadratic, and cubic function of the propensity score P_t $(= e_t(X_t))$, along with their associated β_t coefficients and a final ϵ_t term for the error. Furthermore, M_u analogously models the untreated subgroup outcome. Finally, we apply these models to obtain the values for $\hat{y}_t$ and $\hat{y}_u$ for each individual across the subgroups. Depending on the estimand, we need to pass the propensity scores of either the untreated or treated group to the models, as follows:

$$\hat{y}_t = \begin{cases} M_t(P_t), & \text{for ATT estimand} \\ M_t(P_t), & \text{for ATE estimand} \end{cases} \quad \hat{y}_u = \begin{cases} M_u(P_t), & \text{for ATT estimand} \\ M_u(P_u), & \text{for ATE estimand} \end{cases}$$

The raw mean difference, however, is not generally stable and homogeneous because it depends on the unit of measurement of the effect variable. Therefore, measures have been developed to quantify the effect size in a standardized manner (Ledesma et al. 2009). For treatment effects based on means, the most commonly used ones include Glass's Delta, Hedges's g, and Cohen's d. We have adopted Glass's Delta, which is defined as

$$\text{Glass's } \Delta = (\overline{y}_t - \overline{y}_u)/\sigma(y_u). \tag{14}$$

where $\overline{y}_t$ and $\overline{y}_u$ are the (weighted) mean outcome of the treated group and untreated group, respectively, and $\sigma(y_u)$ is the related standard deviation of the outcome of the untreated group.

B. Software and algorithms

Table 4: List of software packages designed for causal inference (Nogueira et al. 2022).

Language	Name	Description
Python	DoWhy	Causal inference
Python	CausalML	Machine learning, causal inference
Python	EconML	Machine learning, causal inference
R	DoWhy	Causal inference
Python	Matching	Matching
R	MatchIT	Mathing
R	R-FLAME	Matching
Python	dame-flame	Matching
R	PSW	Propensity score
R	ipw	Inverse probability
R	PSweight	Inverse probability
R	RISCA	Causal inference, cohort-based analysis
R	CausalGAM	Inverse propensity scores methods
R	tmle	Targeted maximum likelihood estimator
R	BART	Bayesian Additive Regression Trees
R	grf	Generalized Random Forests
Python	CEVAE	Causal Effect Variational Autoencoder
Python	SITE	Individual Treatment Effect, Deep Representation Learning
R, Python	rdrobust	Regression Discontinuity Design
R	rddtools	Regression Discontinuity Design
R	rdd	Regression Discontinuity Design
R	plm	Panel data
Python	linearmodels	Panel data, instrumental variables
R	Synth	Synthetic Control Method
R	causalimpact	Synthetic Control Method

Table 5: Overview of software packages for causal discovery (graph mining algorithms) in observational data (Nogueira et al. 2022).

Software	**Type of data**				**Type of algorithm**			
	Categorical	Continuous	Mixed	Time-series	Causal sufficiency	Constraint-based	Score-based	Non-Bayesian
bnlearn								
MMPC	•	•	•		•	•		
PC	•	•	•		•	•		
pcalg								
AGES	•	•	•		•		•	
FCI	•	•	•			•		
FCI_JCI	•	•	•			•		
Anytime FCI	•	•	•			•		
Adaptative Anytime FCI	•	•	•			•		
FCI+	•	•	•			•		
GDS	•	•	•		•		•	
GES	•	•	•		•		•	
GIES	•	•	•		•		•	
LINGAM	•	•	•					•
PC	•	•	•		•	•		
CPC	•	•	•		•	•		
PC Select (PC simple)	•	•	•		•	•		
RFCI	•	•	•			•		
Tetrad								
PC and PC-Stable	•	•	•		•	•		
CPC and CPC-Stable	•	•	•		•	•		
PcMax	•	•	•		•	•		
FGES/FGES-MB	•	•	•		•		•	
IMaGES	•	•	•				•	
FCI	•	•	•			•		
RFCI/RFCI-BSC	•	•	•			•		
GFCI	•	•	•				•	
MBFS	•	•	•		•	•		
GLASSO	•	•	•					•
FOFC	•	•	•		•	•		
FTFC	•	•	•					
LiNGAM	•	•	•					•

Table 6: List of software packages for computing effect size estimands, extracted from the work of Nogueira et al. (2022).

Language	**Name**	**Estimand**				
		ATE	ATT	ITE	CATE	LATE
Python	dowhy	•	•	•	•	•
Python	econML	•	•	•	•	•
R	Matching	•	•		•	
R	MatchIT	•	•		•	
R	R-FLAME	•			•	
Python	dame-flame	•			•	
R	PSW	•	•		•	
R	CBPS	•	•		•	•
R	ipw	•	•		•	
R	PSweight	•	•		•	
R	RISCA	•	•			
R	CausalGAM	•	•			
R	tmle	•	•		•	
R	BART	•	•		•	
R	grf	•	•		•	
Python	causalML	•	•	•	•	•
Python	CEVAE	•		•	•	
Python	SITE	•	•		•	
-	ivreg					•
-	rddrobust	•				
R	rddtools	•				
Python	rdd	•				
R	plm	•				
Python	linearmodels	•				
R	Synth	•				
R, Python	causalImpact	•				

C. User feedback

During the experiment we collected the following user feedback.

1. Confounding (and bias) is one of the major problems in current clinical research. Therefore, having a (complementary) support system for this purpose can bring added value.
2. The real-time iterative estimation of the treatment effect is highly appreciated during discussions.
3. Adaptive graph reduction based on the Markov Blanket was helpful.
4. The mined graph was interesting, but also taken with a grain of salt. Consider guiding the mining algorithm using the expert graph.
5. Consider support for importing expert graphs from elsewhere next to mined graphs and show certainty on edges (perhaps also for mined graphs).
6. Add version support to the graph editor, along with annotations and certainties (and use only edges with enough certainty for computations).
7. Add markers for edges with uncertain direction to let the researcher know what to measure and establish in clinical practice to update the graph.

Affiliation:

Dennis Dingen
Department of Mathematics & Computer Science
Eindhoven University of Technology
Eindhoven, The Netherlands
E-mail: d.j.w.g.m.dingen@tue.nl

Marcel van 't Veer
Cardiology department
Catharina Hospital Eindhoven
Eindhoven, The Netherlands
E-mail: marcel.vh.veer@catharinaziekenhuis.nl

Tom Bakkes
Department of Electrical engineering
Eindhoven University of Technology
Eindhoven, The Netherlands
E-mail: t.h.g.f.bakkes@tue.nl

Erik Korsten
Department of Anesthesiology
Catharina Hospital Eindhoven
Eindhoven, The Netherlands
E-mail: erik.korsten@catharinaziekenhuis.nl

Arthur Bouwman
Department of Anesthesiology
Catharina Hospital Eindhoven
Eindhoven, The Netherlands
E-mail: arthur.bouwman@catharinaziekenhuis.nl

Jarke J. van Wijk
Department of Mathematics & Computer Science
Eindhoven University of Technology
Eindhoven, The Netherlands
E-mail: j.j.van.wijk@tue.nl

Journal of Data Science, Statistics, and Visualisation https://jdssv.org/
published by the International Association for Statistical Computing
http://iasc-isi.org/

April 2024, Volume IV, Issue III *Submitted:* 2022-08-05
doi:10.52933/jdssv.v4i3.72 *Accepted:* 2023-08-25