



Counterfactual inference with latent variable and its application in mental health care

Guilherme F. Marchezini¹ · Anisio M. Lacerda¹ · Gisele L. Pappa¹  · Wagner Meira Jr.¹ · Debora Miranda¹ · Marco A. Romano-Silva¹ · Danielle S. Costa¹ · Leandro Malloy Diniz¹

Received: 25 April 2021 / Accepted: 18 December 2021 / Published online: 31 January 2022

© The Author(s), under exclusive licence to Springer Science+Business Media LLC, part of Springer Nature 2022

Abstract

This paper deals with the problem of modeling counterfactual reasoning in scenarios where, apart from the observed endogenous variables, we have a latent variable that affects the outcomes and, consequently, the results of counterfactuals queries. This is a common setup in healthcare problems, including mental health. We propose a new framework where the aforementioned problem is modeled as a multivariate regression and the counterfactual model accounts for both observed and a latent variable, where the latter represents what we call the patient individuality factor (φ). In mental health, focusing on individuals is paramount, as past experiences can change how people see or deal with situations, but individuality cannot be directly measured. To the

Responsible editors: Annalisa Appice, Sergio Escalera, Jose A. Gamez and Heike Trautman.

✉ Gisele L. Pappa
glpappa@dcc.ufmg.br

Guilherme F. Marchezini
guilhermefmar@gmail.com

Anisio M. Lacerda
anisio@dcc.ufmg.br

Wagner Meira Jr.
meira@dcc.ufmg.br

Debora Miranda
Debora.m.miranda@gmail.com

Marco A. Romano-Silva
romanosilva@gmail.com

Danielle S. Costa
daniellesouza.psi@gmail.com

Leandro Malloy Diniz
Malloy.diniz@gmail.com

¹ Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil

best of our knowledge, this is the first counterfactual approach that considers *both observational and latent variables* to provide *deterministic* answers to counterfactual queries, such as: what if I change the social support of a patient, to what extent can I change his/her anxiety? The framework combines concepts from deep representation learning and causal inference to infer the value of φ and capture both *non-linear and multiplicative effects* of causal variables. Experiments are performed with both synthetic and real-world datasets, where we predict how changes in people's actions may lead to different outcomes in terms of symptoms of mental illness and quality of life. Results show the model learns the individually factor with errors lower than 0.05 and answers counterfactual queries that are supported by the medical literature. The model has the potential to recommend small changes in people's lives that may completely change their relationship with mental illness.

Keywords Counterfactual inference · Multivariate regression · Mental health

1 Introduction

What if? questions are a common follow-up of the decision making process in almost all domains of knowledge. For example, a physician recommends treatment *A* to a patient, but her health condition worsens and she dies. Then a typical question is: What would have happened if treatment *B* had been followed? These types of questions may be answered with counterfactual inference of hypothetical scenarios (Pearl et al. 2016).

In counterfactual inference, one models alternative scenarios that are fictional, once it is not possible to change what has already happened. Let us take the patient who received treatment *A* and the event outcome of her death (*d*), where our evidence is $E = d$. Counterfactual inference can compute either a probability or a deterministic value of a different outcome (evidence), e.g., the patient survives ($E = s$), counter to the fact that she died—if a certain precondition in the given scenario was different, e.g., she had received treatment *B*. In contrast with traditional machine learning approaches, counterfactual inference involves both observational data and assumptions about the world, being a powerful framework for answering questions such as the one posed above.

The *structural causal model* (SCM) framework is the standard approach for calculating the outcome of counterfactual queries (Pearl 2009). An SCM consists of two sets of variables U and V , and a set of functions f that assigns each variable in V a value according to the values of the other variables in the model. The variables in U are called exogenous, in the sense that they are external to the model and we chose not to explain how they are caused. The variables in V are endogenous and each of them is a descendant of at least one exogenous variable. By knowing the values of the exogenous variables, one can use the functions f to accurately determine the value of every endogenous variable. The SCM is often associated with a graphical causal model, which represents the causal relations of the endogenous variables and where each exogenous variable is a root in the graph.

Counterfactual inference has been explored in healthcare applications in the past (Pawlowski et al. 2020), especially in the problem of estimating the individual treat-

ment effect (ITE) (Zhang and Bareinboim 2020; Shalit et al. 2017; Bica et al. 2020). This scenario is equivalent to the one illustrated above, and, in this case, the observational data (endogenous variables) contains past actions (medications), their outcomes (evidences), additional context information (records of patients), but no access to the mechanisms that led to an action (why each medication was given to each patient). Counterfactual inference is specially used to try to remove bias or other confound factors from models. Usually, this is necessary as patients in worse health conditions receive a different type of treatment than patients who are in a better health condition. Counterfactual inference is then simulated using a less biased model.

In this paper, we deal with the problem of modeling counterfactual reasoning in scenarios where, apart from the observed endogenous variables, we have a latent variable that affects the outcomes and, consequently, the results of counterfactuals queries (or simply counterfactuals). In this case, a single deterministic model can make different predictions to input instances that are the same, because their prediction values vary due to the latent variable. For example, in mental health, focusing on individuals is essential, as past experiences may change how people see or deal with situations (Friedman et al. 2014). However, one's individuality cannot be directly measured, as it is inherent to each person, and may be considered a latent variable in many scenarios where individuality impacts outcomes, e.g., levels of stress in the case of mental health or performance in the case of an online course.

The existence of latent variables is common in real-world observational studies since we cannot directly measure variables such as personal preferences or most genetic, psychological, and environmental factors. In studies with individuals, we can use proxy variables to deal with model confounders (Montgomery et al. 2000; Louizos et al. 2017). For example, when a patient receives treatment for a disease, we may have a proxy variable, e.g., the results of an exam, to indicate how ill he already was before receiving a specific type of treatment.

We evaluate the proposed counterfactual model by analyzing how the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) pandemic is affecting peoples' mental health. Our focus is not on treatment (and hence our observational data does not have actions), but on how changing one's behavior (which represents our observed variables) may affect the quality of life - measured using the World Health Organization Instrument to evaluate Quality of Life (WHOQOL-BREF)¹ and symptoms related to conditions such as anxiety and depression, measured by the Brief Symptom Inventory (BSI) instrument (Derogatis 1993) (our outcomes). The resulting model is able to answer questions such as: "what if I increase the support from friends of a person, to what extent can I improve his/her quality of life and decrease anxiety levels?".

Both WHOQOL and BSI are commonly used instruments for assessing mental health. They are built based on questionnaires, where patients answer a set of questions related to their mental health and quality of life. One major difference of this setting, when compared to other counterfactual scenarios, is that we assume there is a common latent variable that affects both outcomes—WHOQOL and BSI—simultaneously (see Fig. 2), and we name it the individuality factor of a patient. However, as previously mentioned, the individuality factor cannot be directly measured.

¹ <https://www.who.int/tools/whoqol>.

Personalized mental health treatment has the potential to be more effective than traditional clinical and statistical approaches, where the goal is often to identify alternatives to treatments that explain the benefits and variance for the “*majority of a clinical group*”, and are formally tested for “*group effects*” (Bzdok and Meyer-Lindenberg 2018). Further, one may also target adjustment to treatments by identifying both (i) the most appropriate set of clinical interventions and/or (ii) their form of delivery related to a particular individual. Finally, properly understanding causal effects of treatments for individuals may enable more effective just-in-time adaptive interventions, which tend to create more engaging and adaptive treatments (Nahum-Shani et al. 2018).

In this setting, we propose a new framework that models our problem as a multivariate regression, where the counterfactual model considers we have both observed variables and a latent variable φ , which represents the individuality factor. Note that φ is not exogenous, as it is considered in the model, but it is not observed. A multivariate regression is used because, by definition, φ impacts more than one outcome variable in the scenarios we are dealing with—in the case of mental health, WHOQOL and BSI. To the best of our knowledge, this is the first counterfactual approach that considers *both observational and latent variables* (see Table 1). The framework combines concepts from deep representation learning and causal inference to infer the value of φ and provide *deterministic* answers to counterfactual queries—in contrast to most counterfactual models that return probabilistic answers. As we are dealing with individuals, deterministic methods are preferred over probabilistic. The proposed model can go beyond the *linear and additive effects* and can also capture both *non-linear and multiplicative effects* of causal variables. Therefore, we can use in the context of mental health without assuming how they would manifest.

We perform experiments with synthetic datasets to validate our assumptions and approach, and then used data obtained from a mental health questionnaire answered by 153,514 subjects in the year 2020 to assess the environmental, physical, and psychological effects of the pandemics in their lives. For the mental health scenario, we model the causal variables relating patients and outcomes, enabling clinical interventions as counterfactual queries. As evaluating counterfactual approaches is a challenge, we had specialists to analyze the counterfactual queries, which demonstrated the effectiveness of our proposed approach.

In summary, the main contributions of this paper are as follows:

- We define a new problem setting to compute counterfactual queries in multivariate regression scenarios, where causal questions can be answered considering both observational and latent data.
- We present a new deep neural network to compute counterfactual queries at populational and individual levels.
- We empirically evaluate the framework in two types of data: (i) a set of synthetic simulated datasets and (ii) a real-world dataset built over questionnaires about mental health issues during the COVID-19 pandemic. The results were analyzed by specialists and are supported by the medical literature.

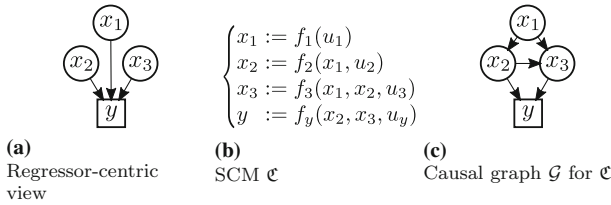


Fig. 1 A view frequently adopted in regular regression models (a) assumes variables are independently manipulated inputs to a given fixed and deterministic regressor h . In the causal approach to counterfactual inference taken in this work, we rather view variables as causally related to each other by a structural causal model (SCM) \mathcal{C} (b) with associated causal graph \mathcal{G} (c)

2 Background

When dealing with causal inference, Pearl (2009) introduced a hierarchy of three levels of problems of increasing power to answer causal questions: (i) association, when association is used in the context of traditional machine learning and we ask questions about the system as it currently is by ignoring causal relationships among features; (ii) intervention, when we act on the system and change its distribution, i.e. we are able to answer population related questions; and (iii) counterfactual, where we can simulate a “parallel world”, i.e., we are able to consider alternative scenarios at individual level. Note that while on the intervention level we can answer questions regarding the system after intervention, on the counterfactual level we can change the system before the intervention.

Counterfactual queries, according to Pearl’s hierarchy, can only be answered by a structural causal model (Pearl 2009).

Definition 1 (Structural Causal Model (SCM)) A structural causal model \mathcal{C} is a 4-tuple $\langle \mathbf{U}, \mathbf{V}, \mathcal{F}, P(\mathbf{U}) \rangle$, where (i) \mathbf{U} is a set $\{U_1, \dots, U_d\}$ of background variables (aka exogenous variables) that are determined by factors outside the model; (ii) \mathbf{V} is a set $\{V_1, \dots, V_d\}$ of endogenous variables, which are determined by other variables in the model, i.e., by variables in $\mathbf{U} \cup \mathbf{V}$; (iii) \mathcal{F} is a set of functions $\{f_1, \dots, f_d\}$ (aka causal mechanisms) that represents structural assignments $v_i := f_i(pa_i, u_i)$, where pa_i is the set of parents of v_i (its *direct causes*); and (iv) $P(\mathbf{U})$ is a probability function defined over the domain of \mathbf{U} .

An SCM is often represented by the associated directed acyclic graph (DAG) \mathcal{G} , which is called the *causal graph* induced by \mathcal{C} . The causal graph is obtained by drawing a directed edge from causes to effects, i.e., from each node in pa_i to v_i for $i \in \{1, \dots, d\}$, see Fig. 1b, c for an example. Since we assume \mathcal{G} is acyclic, every SCM \mathcal{C} implies a unique observational distribution $P_{\mathcal{C}}(\mathbf{X})$, which factorises over \mathcal{G} . Note this satisfies the causal Markov assumption: each variable is independent of its non-effects given its direct causes, i.e. $P_{\mathcal{C}}(\mathbf{V}) = \prod_{i=1}^d P_{\mathcal{C}}(v_i | pa_i)$.

A crucial difference from conventional Bayesian networks is that the conditional factors above imply causal relationships rather than dependence assumptions. Hence, the SCM framework also enables the prediction of the effects of *interventions*, which refers to a situation where some variables are manipulated externally. An interven-

tion that fixes the constant a to x_i is denoted by $\text{do}(x_i := a)$ and called an atomic intervention.

Questions related to interventions and counterfactual layers in the causal hierarchy are defined and computed in terms of the do -operator. However, an important difference among them is that interventions operate at the population level—providing aggregate statistics about the effects of actions, while counterfactuals are queries at the unit level.² Note that the structural assignments (*causal mechanisms*) are changed, but the exogenous variables are the same for the unit. Following, we formally define counterfactuals.

Definition 2 (*Unit-level Counterfactuals* (Pearl 2009)) Let \mathcal{C} be a structural causal model and $\mathcal{C}_{\mathbf{X};\text{do}(a)}$ a modified version of \mathcal{C} , with the equation(s) of X replaced by $X = a$. Denote the solution for Y in the equations of $\mathcal{C}_{\mathbf{X};\text{do}(a)}$ by symbol $Y_{\mathcal{C}_{\mathbf{X};\text{do}(a)}}(u)$. The counterfactual $Y_a(u)$ is given by $Y_x(u) \triangleq Y_{\mathcal{C}_{\mathbf{X};\text{do}(a)}}$.

From this definition, we may effectively provide explanations of the data, given that we are able to manipulate each variable and observe the outcome. The following theorem formalizes the three-step procedure to compute *deterministic* counterfactual queries:³

Theorem 1 (Counterfactual Computation (Pearl 2009)) *Given an SCM \mathcal{C} and evidence e , the counterfactual can be evaluated using the three following steps:*

1. *Abduction* Use evidence $E = e$ to determine the value of U .
2. *Action* Modify the SCM \mathcal{C} , by removing the structural equations for the variables in X and replacing them with the appropriate functions $X = a$, to obtain the modified model $\mathcal{C}_{\mathbf{X};\text{do}(a)}$.
3. *Prediction* Use the modified model $\mathcal{C}_{\mathbf{X};\text{do}(a)}$, and the value of U to compute the value of Y , the consequence of the counterfactual.

Lastly, as formalized in Definition 2, we are interested in deterministic counterfactuals, which refer to hypothetical questions about a single unit of the population, e.g., a patient.

3 Related work

While causal inference concepts have evolved using statistical concepts and the canonical Pearl's definition, these techniques have been recently introduced to deep learning (Schölkopf 2019). For instance, there are investigations in disentangled representation (Locatello et al. 2019), causal reinforcement learning (Zhang and Bareinboim 2020), recourse (Karimi et al. 2021), and model explanation methods (Le et al. 2020; Miller 2020; Guidotti et al. 2018; Wachter et al. 2018).

In principle, causality for deep learning models may provide means to answer interventional and counterfactual questions, although current methods on deep-based

² In the causal inference literature, a *unit* stands for an observed datum, such as an individual patient, an experimental subject, or an agricultural plot Pearl (2009).

³ We refer readers to Theorem 7.1.7 in Pearl (2009) for the probabilistic computation of counterfactuals.

Table 1 Comparison of your work with other from the literature

	Three levels of causation	Tabular data	Latent vars.	Non-linear
Bica et al. (2020)		✓		
Graham et al. (2019)				✓
Pawlowski et al. (2020)	✓			✓
Zhang and Bareinboim (2020)		✓		✓
Zhang et al. (2020)		✓		✓
Ours	✓	✓	✓	✓

counterfactuals are limited by modeling only cause-effect relationships (Singla et al. 2020), or by not providing computation of deterministic counterfactuals (Pawlowski et al. 2020).

The most common approach to causal inference at the unit level refers to estimating the individual treatment effect (ITE) (Zhang et al. 2020; Shalit et al. 2017). The ITE estimation assumes observational data about the individual's context, outcomes, and binary interventions (aka treatments), which results in the observed outcome. A common contribution of these approaches is the analysis of synthetic or semi-synthetic data to validate theoretical analysis of generalization-error bounds. In contrast to these methods, we assume only the individual's context and outcomes are available and focus on learning other latent features that affect the outcomes.

It is worth noting that causal inference techniques have been also applied to sequential data. For instance, Zhang and Bareinboim (2020) consider the dynamic treatment regime, which consists of a sequence of decision rules to determine the most appropriate treatment for patients. They propose a reinforcement learning setting of optimal dynamic treatment regimes provided by the causal diagram of the underlying, unknown environment. Bica et al. (2020) also propose a factor model to predict the treatment effects over time. As we use data from a questionnaire, we do not assume a temporal relation among variables. Additionally, we assume clinical interventions, i.e., intentional action designed to result in an outcome, in the patient's variables to compute counterfactuals, rather than on possible medical treatments for a given disease.

In the context of mental health, Paredes et al. (2014) dealt with personalized clinical interventions. They focused on stress treatment, and made recommendations based solely on observational data. As a result, they were unable to consider causal interventions. We propose to model the causal variables relating patients and outcomes, enabling interventions as counterfactual queries. Additionally, rather than focusing on a specific illness (i.e., stress), our framework is general and enables personalized interventions in both mental health and other well-being dimensions.

The closest work to ours is Pawlowski et al. (2020), which proposes a model for counterfactual inference in high-dimensional data. Their approach focuses on probabilistic counterfactuals (in contrast to ours that focus on deterministic counterfactuals) and high-dimensional data (i.e. images). Similar to ours, they present counterfactual results for individual patients considering brain image data. Here we analyze hypothetical clinical interventions on mental health tabular data.

An important aspect of counterfactual inference is the ability to consider questions at the unit level. Hence, it is important to account for the specific characteristics of individuals, which may be observed or latent. The existence of latent variables is common in real-world observational studies since we cannot measure variables such as personal preferences or most genetic, psychological, and environmental factors. These variables can directly affect the inference of causal relationships from observational data due to confounding. A confounder is a variable which affects both the intervention and the outcome. When a confounder is hidden or unmeasured, we cannot estimate the effect of the intervention on the outcome (Pearl 2009). For example, the socio-economic status of a patient may affect both his access to medication (intervention) but also his general health (outcome). In this case, socio-economic status is a confounder between medication and health outcomes.

Most works deal with confounders by using proxy variables, e.g., in the case above, one can use the zip code as a proxy to socio-economic status, since they are correlated (Miao et al. 2018; Montgomery et al. 2000; Wooldridge 2009). One of the exceptions to this approach is the work of Louizos et al. (2017), where the authors use a variational autoencoder to simultaneously discover the hidden confounders and infer how they affect treatment and outcome.

However, note that proxy variables are used to obtain the maximum possible available information to predict an outcome. In this paper, we focus on counterfactual analysis and queries. We use past information and the outcome to infer latent information.

Furthermore, our latent variable can affect more than one outcome, as the individuality factor transverses many aspects of an individual. As far as we know, we are the first work proposing to model latent variables that affect multiple outcomes. As noted in Bzdok and Meyer-Lindenberg (2018), several psychopathological symptoms in psychiatric patients are shared among psychiatric disorders, rather than being unique to them. As we model the presence of these latent variables, we are able to answer questions on the third level of the hierarchy of causation—structural counterfactual.

Table 1 summarizes how previously works fulfill the requirements of our framework. Within the domain of health, there is one work that can reach the three levels of hierarchy, as we can in this work. However, other papers in the literature are able to “simulate” the counterfactual level, as the SCM is implicit to the problem. We say they “simulate” it because they assume they already know all data necessary to perform the inference, which is a non-realistic scenario (e.g., Bica et al. (2020), Graham et al. (2019)).

4 Computing structural counterfactuals

In this section, we introduce a new framework that combines concepts of deep learning representation and structural causal models to (i) infer associations among dependent and independent variables (level 1 of causal hierarchy); (ii) find values for the latent individuality factor; (iii) enable interventions in the population level (level 2 of causal hierarchy); and (iv) compute unit-level counterfactuals (level 3 of causal hierarchy).

Fig. 2 Causal DAG with multiple outcome y^k , features x , and latent individuality factor φ . Calculated values appear in black boxes, observed variables in black circles, and unobserved variables in white

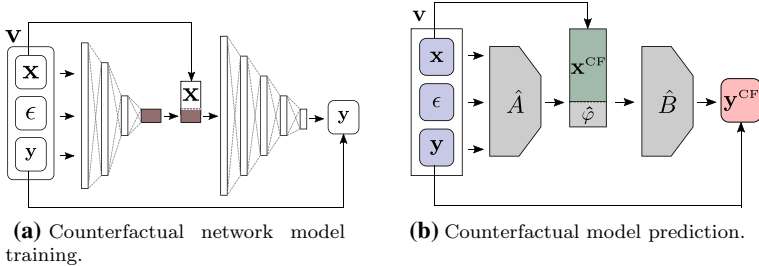
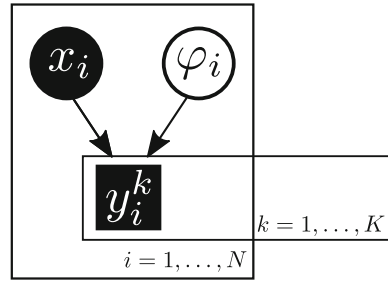


Fig. 3 An illustration of the proposed counterfactual model, highlighting the factual (i.e. $\mathbf{v} = \langle \mathbf{x}, \epsilon, \mathbf{y} \rangle$) and counterfactual inputs (i.e. \mathbf{x}^{CF}) to compute the factual (i.e. \mathbf{y}) and counterfactual outcome (i.e. \mathbf{y}^{CF})

4.1 Problem setting

We define our problem as computing deterministic structural counterfactuals in multivariate regression scenarios. We work with observation data, where we have the response vectors (the observed effects) $\mathbf{y}_i = (y_{i1}, \dots, y_{ik}) \in \mathbb{R}^k$, and the predictor vectors (the causes) $\mathbf{x}_i = (x_{i1}, \dots, x_{id}) \in \mathbb{R}^d$. A crucial property of our model is that it assumes the presence of an individuality factor that affects the response vectors and which is both latent and immutable. We model the individuality factor of each instance as latent vectors (non-observed causes) $\varphi_i = (\varphi_{i1}, \dots, \varphi_{il}) \in \mathbb{R}^l$. The DAG \mathcal{G} of the resulting SCM is presented in Fig. 2. From the graph, we can learn the causal mechanisms to estimate the impact of populational and counterfactual interventions (Pearl 2009). Following the convention of Buesing et al. (2018), calculated values are shown in black boxes, observed variables in black circles, and unobserved variables in white.

The counterfactual queries are defined at the unit level, where the causal mechanisms are changed, while the exogenous and latent variables are kept the same for the observed instance. Assuming a structural causal model \mathcal{C} that factorizes as a DAG, we specify the causal mechanisms F using a deep neural network. We also assume the full specification of \mathcal{F} and \mathcal{F}^{-1} , such that $\mathcal{F}(\mathcal{F}^{-1}(\mathbf{x})) = \mathbf{x}$. Hence, one can determine the distinct values of individuality factors that give rise to a particular realization of the endogenous variables, $\{X_i = \mathbf{x}_i^F\}_{i=1}^N$, as $\mathcal{F}^{-1}(\mathbf{x}^F)$. To compute the counterfactual queries, we assume a set of actions, A , in a world \mathcal{C} , which yields to an updated world model \mathcal{C}_A with structural equations $F_A = \{F_i\}_{i \neq I} \cup \{X_i := a_i\}_{i \in I}$. As a consequence, we can compute any structural counterfactual queries \mathbf{x}^{SCM} , which automatically accounts for inter-variable causal dependencies, for an instance \mathbf{x}_A^F as

$\mathbf{x}^{SCF} = \mathcal{F}(\mathcal{F}^{-1}(\mathbf{x}^F))$, read as: “given the model \mathcal{C} and having observed \mathbf{x}^F , what is the value of the endogenous variables if the set of actions A is performed?”.

Summarizing, we make the following assumptions:

- (i) There are no other causes of Y except the latent individuality factor φ , i.e., we assume *determinism*: a set of functions f_y exists such that $Y = f_y(X, \varphi)$.
- (ii) X and U are independent: $X \perp\!\!\!\perp \varphi$.
- (iii) φ variables are unchanged by hypothetical counterfactual interventions, i.e. they are immutable regardless of being latent.

4.2 Our framework

The computation of our counterfactual framework is divided into three phases: (i) from the observational data, we apply a *causal graph discovery algorithm* to generate the causal graph, which will determine which predictor variables have a causal effect on the response variables; (ii) we run a *base model h* that works as an interpolator, and learns associations between the predictor and response variables with an error close to 0; (iii) we propose the individuality factor discovery counterfactual network, *ICoNet*, which finds the latent individuality factors φ shared by the output variables. It does that by using the residuals of the base model together with the predictor and response variables. Having the values of φ , we can perform counterfactual predictions by making interventions to the variables of interest.

The main rationale of the proposed method is that the residuals of the interpolator, detailed in Sect. 4.2.2, correspond to the effects of the latent variables that directly affect Y^F . Hence, we propose a deep neural network that learns the values of these latent variables by using both the predictors, response variables, and residuals, as detailed in Sect. 4.2.3.

4.2.1 Causal graph discovery

The causal relations among predictor and response variables are not artifacts that can be directly read from data alone and need to be inferred. But to compute counterfactual queries, we need to explicitly state the causal relations among the variables. However, the real causal graph is unknown in most practical applications, and learning directed acyclic graphs from observational data is an NP-hard problem (Chickering 1996; Chickering et al. 2004).

The field of causal discovery is concerned with the problem of inference of causal relations (Eberhardt 2017; Zheng et al. 2018). Causal discovery algorithms have been successfully applied in practical health-related problems. For instance, Shen et al. (2020) considered data about Alzheimer’s disease and focused on investigating whether causal discovery frameworks were able to discover known causal relationships from observational clinical data. They found that causal discovery frameworks should be used with longitudinal data providing full prior knowledge available. We refer the readers to Tennant et al. (2019), which survey the use of causal DAGs applied to health care.

Here we use the framework NO TEARS (Zheng et al. 2018) to find the causal graph, which formulates the graph discovery learning problem as a continuous constrained optimization task, leveraging an algebraic characterization of DAGs. More formally, NO TEARS solves the following continuous optimization problem:

$$\begin{aligned} \min_{\mathbf{G} \in \mathbb{R}^{d \times d}} \quad & \frac{1}{2N} \|\mathbf{X} - \mathbf{X}\mathbf{G}^T\|_2^2 + \lambda \|\mathbf{G}\|_1 \\ \text{s.t.} \quad & h(\mathbf{G}) = 0, \end{aligned}$$

where \mathbf{X} is the input data, the function $h(\cdot) : \mathbb{R}^{d \times d} \rightarrow \mathbb{R}$ is continuous, such that $h(\mathbf{G}) = 0$ iff the graph is acyclic, and $\mathbf{G} = (\mathbf{g}_1, \dots, \mathbf{g}_d)'$ is the weighted adjacency matrix.

4.2.2 Base model

A crucial component of our framework is that we assume a *base model*, or *interpolator*, which is an estimator that achieves almost zero training error. We could have used any interpolator, but we opt for a multilayer perceptron (MLP) network trained with the L_1 norm as the loss function and the ADAM optimizer.

We adopted such strategy because, although modern deep learning models involve several parameters, deep neural networks are frequently observed to generalize well in practice (Zhang et al. 2016). The conventional wisdom is that interpolation (vanishing training error) leads to overfitting or poor generalization. However, it has been recently shown that these concepts are in general distinct, and interpolation does not contradict generalization (Belkin et al. 2019).

In particular, Belkin et al. (2019) discuss the shape of risk curves in scenarios of modern high-complexity learners, such as Deep Neural Networks. Assuming n as a fixed training sample size, they show that risk curves represented as a function of some (approximate) measure of its complexity N (e.g. the number of features) can display what they call double descent. By increasing N (i.e. complexity), the risk initially decreases, attains a minimum, and then increases until N equals n . This is the state of the model that we are interested in, where the training data are fitted perfectly (i.e., overfitted). According to Belkin et al. (2019), increasing N even further leads to risk decreasing a second and final time, creating a peak at $N = n$. In our work, we aimed to train our network to stabilize on this peak.

This double descent behaviour is clearly related to one of the important tenets of the machine learning field, the bias-variance trade-off, which implies that a model should balance underfitting and overfitting. The model needs to be complex enough to express underlying structure in data and simple enough to avoid fitting spurious correlations (i.e. noise). However, in modern literature, very complex models such as neural networks are trained to exactly fit (i.e., interpolate) the data. In a classical view, these models are overfitted, and yet they often reach high accuracy on test data, which is our case. Belkin et al. (2019) showed that this apparent contradiction can be explained by assuming the double descent for learning curves of machine learning models. Actually, the double descent phenomenon has been studied before.

For instance, in Vallet et al. (1989), the authors empirically demonstrated double descent for classifiers through minimum norm linear regression.

The double descent behavior is an essential and strong assumption for the conception of our model. We exploit the interpolation property of neural networks to isolate the individuality factor φ shared by our outcome variables. Since we can achieve almost zero error in training, the residual error can be attributed to the latent individuality of each unit.

Hypothetically, assume our model is represented by $y = f(x) + \epsilon$, where $f(x) = \mu(x) + \Phi(x)$. $\mu(x)$ is the part of the model that is common to everyone, for example, the effect a treatment has on patients in general. $\Phi(x)$ is what we call the individuality factor, which changes the effect of the medicine in a specific patient. When we overfit the model, we have y with an almost 0 error. Our current model does not separated $\mu(x) + \Phi(x)$, and this is certainly a very interesting direction for future work.

Note that we train a different base model for each target variable in our multivariate regression setting. Hence, for each of our k target variables, we compute the training error $\epsilon_{\mathbf{k}} = \mathbf{y}_{\mathbf{k}} - \hat{\mathbf{y}}_{\mathbf{k}}$, where $\hat{\mathbf{y}}_{\mathbf{k}}$ stands for the prediction of the base model.

4.2.3 Individuality factor discover counterfactual network

We propose a new network architecture to compute structural counterfactual queries based on the individuality factor, named Individuality factor discover Counterfactual Network (ICoNet). Figure 3a depicts the main components of the architecture. The network receives as input a vector \mathbf{v} , which includes (i) the independent variables \mathbf{x} , (ii) the dependent variables \mathbf{y} , and (iii) the error $\epsilon = \mathbf{y} - \hat{\mathbf{y}}$, where $\hat{\mathbf{y}}$ refers to base model prediction. ICoNet is trained to learn the output \mathbf{y} using the L1-distance as the loss function. We model the network architecture in a way that we create two bottlenecks. The first forces an intermediate layer to learn the variables independent from X that are necessary to correctly predict Y . This variable, in our case, is the individuality factor φ —represented by a single neuron shown by a full brown square in Fig. 3a. We can learn φ when we give the observed variables \mathbf{X} as input to a merger layer that comes right after the end of the first bottleneck, which has a single neuron. The merger layer is followed by a sequence of feed-forward layers, which will form a second bottleneck.

It is important to note that this architecture is capable of predicting φ at the end of the first bottleneck because we deal with a multivariate regression problem (we predict more than one output) and map it to a single neuron. This same architecture will not work in cases of simple multivariate regression (where a single y needs to be predicted) or when the number of neurons at the end of the bottleneck is greater than the number of variables y to be predicted. If that is the case, the network will simply learn to predict y .

The first bottleneck has three layers with Relu activation, with 40, 20, and 5 units followed by one linear unit. We use the Relu as we may deal with negative inputs. They are followed by the merge layer with 100 units, which concatenates the final unit with the X input. The second bottleneck has 4 layers with Relu activation and 500, 500, 250, and 50 units, respectively. The output is a sigmoid layer with two neurons (one for each output), as the datasets have their outputs normalized in $[0, 1]$. The network was

trained with ADAM. The number of neurons in each layer and the number of layers were determined in preliminary experiments, where we tested the bottleneck concept while prioritizing speed of convergence. As the main objective was to reach the double descent property, and the proposed configuration led to that, further experiments on different architectures and hyperparameters were left for future work.

Model training Formally, we define the counterfactual neural network as the solution of an optimization problem considering the t-tuple $\langle \mathcal{A}, \mathcal{B}, \mathbf{V}, \Delta \rangle$ where:

1. \mathcal{A} is a class of functions from \mathbb{R}^{d+2k} to \mathbb{R}^{d+1} .
2. \mathcal{B} is a class of functions from \mathbb{R}^{d+1} to \mathbb{R}^{d+2k} .
3. \mathbf{V} input vector comprises the following three vectors: (i) \mathbf{X} : a set of n training vectors in \mathbb{R}^d ; (ii) \mathbf{Y} : the corresponding set of n target vectors in \mathbb{R}^k ; and (iii) \mathbf{E} : the set of n base model errors in \mathbb{R}^k .
4. Δ is a dissimilarity or distortion function defined over \mathbb{R}^{d+2k} .

For any $A \in \mathcal{A}$ and $B \in \mathcal{B}$, the counterfactual network transforms an input $\mathbf{v} \in \mathbb{R}^{d+2k}$ into an output vector $A \circ B(\mathbf{v})$. The model training refers to find $A \in \mathcal{A}$ and $B \in \mathcal{B}$ that minimize the overall distortion function:

$$\min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^n \mathbb{E}(\mathbf{v}_i, x_i, \mathbf{y}_i) = \min_{\mathbf{A}, \mathbf{B}} \sum_{i=1}^n \Delta(A \cdot p[B(v_i), x_i], y_i) \quad (1)$$

Counterfactual prediction After the model is trained, it can be used to perform counterfactual predictions. Figure 3b illustrates this process, where the counterfactual prediction depends on the factual instances and previously learned counterfactual network parameters. The network receives as input the same factual inputs \mathbf{x} , targets \mathbf{y} , and errors ϵ . Additionally, the model is given the hypothetical or counterfactual values of instance features, i.e., \mathbf{x}^{CF} , to the merge layer. Given the previous inputs, the structural model (i.e., ICoNet) can learn the target counterfactuals \mathbf{y}^{CF} .

5 Experimental analysis

This section shows experimental results in 53 synthetic datasets and a case study on mental health during the COVID pandemic. In all experiments, datasets were divided into training (70%) and test (30%) sets. As our main objective is not to make factual predictions, we did not perform any cross-validation. Experiments were run on one GPU (GEFORCE 1060 6GB), and the framework was implemented in Python 3.6.12 and Keras 2.4.3. The code and more implementation details are available online.⁴

Note that we did not find a suitable baseline from the related work, as they would require a complete remodeling to be adapted to this problem. The methods proposed by Bica et al. (2020), Zhang and Bareinboim (2020), and Zhang et al. (2020) focus on ITE, and consider the bias between the patient situation, the treatment and the result. They were modeled to query counterfactual in the treatment variable, not in the input variables as we do. Pawlowski et al. (2020) focused in probabilistic counterfactual and

⁴ <https://github.com/marchezinixd/ICoNet>.

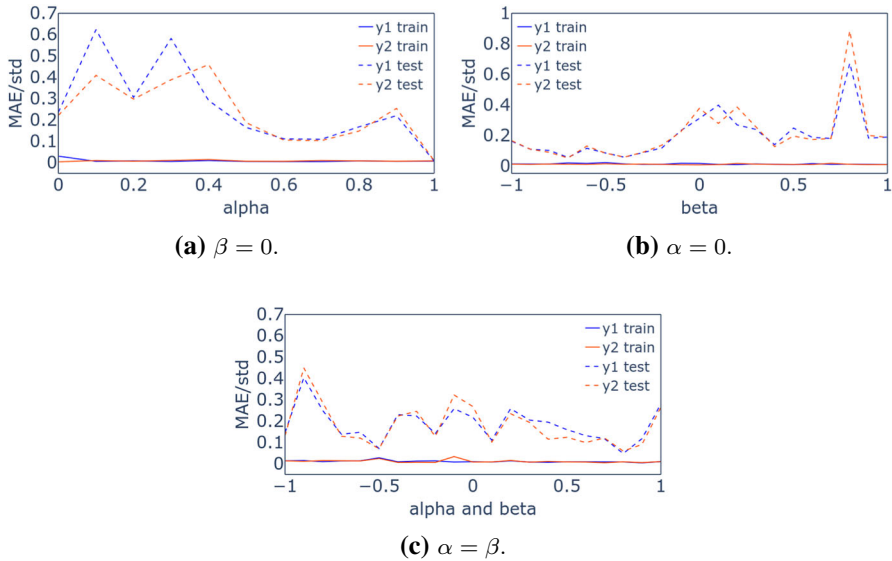


Fig. 4 Comparison of train and test *normalized association errors* for different combinations of α and β

explicitly assumed the correctness of the model only in high dimensional data, and we work with tabular data. Finally, the method proposed by (Pearl 2009) can solve linear problems with additive noise, which is a characteristics of three of our 53 synthetic datasets.

The base models follow the same architecture for the synthetic and real data except for the last unit, which has a linear activation function for the synthetic data and a sigmoid for the case study, as the values of y are normalized in $[0,1]$. The base network for the synthetic data was an MLP with 5 tanh layers of 50 units, trained for 250 epochs with a 0.001 learning rate, the Adam optimizer and mini-batches of size 128. Since we are searching for a double descent, we did not use the validation step for the early stopping. The counterfactual network, ICoNet, was trained for 500 epochs, a learning rate of 0.001 and mini-batches of size 128.

5.1 Experiments with synthetic data

In this first experiment, we generate synthetic data for both X and $Y = \{y_1, y_2\}$ and the individuality factor φ , which is latent in our real-world dataset. The idea of this experiment is to show the proposed model can generate the individuality factor and then answer counterfactual queries. It also shows the model deals in different ways with linear and non-linear problems.

The synthetic data was generated using a function borrowed from the experimental setup followed by Stegle et al. (2010). We generated 53 simulated datasets using the model $Y = (X + \beta X^3)e^{\alpha E} + (1 - \alpha)E$, where the random variable X was sampled from a normal distribution with mean 0 and variance 1.

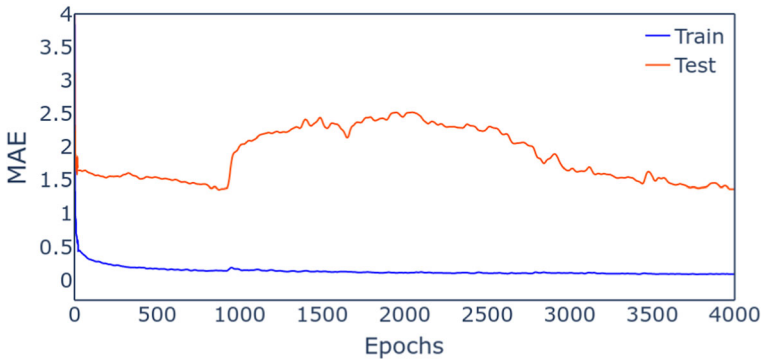


Fig. 5 Double descent behavior of the base model in the test loss

We consider E as our individuality factor φ , since it is independent of X and affects Y . This is an interesting model to use because the parameter α controls the type of effect the observed individuality factor φ has on Y , interpolating between purely additive noise ($\alpha = 0$) and pure multiplicative noise ($\alpha = 1$). The coefficient β , in turn, determines the non-linearity of the true causal model, with $\beta = 0$ corresponding to the linear case.

The 53 independent datasets were generated considering different values of α and β . Each dataset has $N = 10,000$ samples, and were generated following three different scenarios. In the first, $\beta = 0$ and α ranges from 0 to 1 with a step of 0.1. In the second, $\alpha = 0$ and β ranges from -1 to 1 in steps of 0.1. In the last, $\alpha = \beta$ as values range from -1 to 1, with steps of 0.1. In all cases, an equal number of instances was assigned values of E equals 1, 2, and 3. The two values of Y were generated using the same equation, but with different values for X .

Recall that the first step of the proposed method is the causal graph discovery, which here is unnecessary as we have a single variable that we know affects y . The next step is to train the base model, which will generate an error close to 0. Following we run ICoNet to find the individuality factor.

Figure 5 shows that, when running the proposed method in the synthetic dataset (a similar behavior is found for the real-world mental health dataset), both training and test error decrease as a function of model complexity, measured in terms of epochs, and consequently, parameter convergence. This behavior follows a *double descent* curve, which is a mark of overparameterized interpolator models and the assumption we rely on for the base model to work.

Following, Fig. 4 shows the values of normalized mean absolute error (MAE) (MAE/standard deviation) of y_1 and y_2 for the factual prediction of ICoNet, considering the three scenarios used to generate the synthetic datasets. In Fig. 4, observe that the training error is really low for the three scenarios, but the test error differs significantly. In Fig. 4a the test error starts high when α is low and decreases as α increases and the individuality factor adds a multiplicative effect in Y . In Fig. 4b, as β gets closer to 0 (the linear case) the error increases, meaning that the model is learning to generalize better for non-linear cases. In Fig. 4c we observe that the test error was

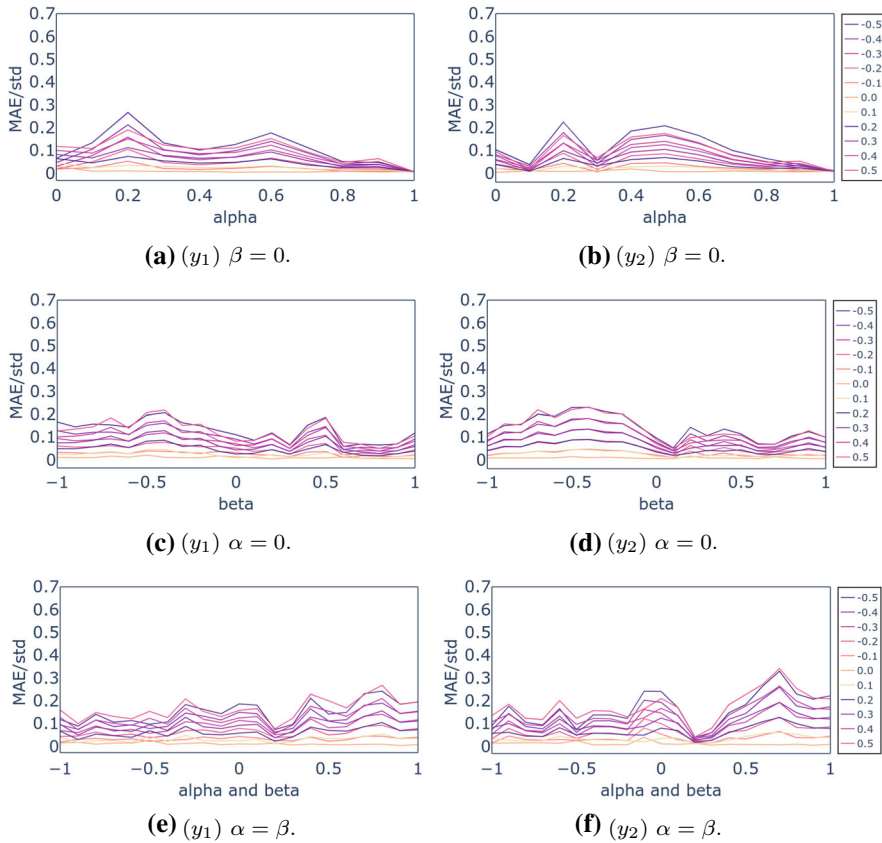


Fig. 6 Comparison of *normalized counterfactual errors* for simulated data with different combinations of α and β : the left column shows the values for y_1 and the right column for y_2

kept below 0.4 in almost all series, we can see a little peak when β gets closer to 0, but the measure is lower than 0.3, which is considered a low value. In sum, these results show that the model has good results with non-linear and multiplicative cases that usually are the characteristics of real-world models.

Next, we run ICoNet to evaluate the counterfactual queries. They were generated for all samples in the training set for both Y_1 and Y_2 with the formula $X^{cf} = X + q$, where q values ranged from -0.5 to 0.5 in 0.1 steps. As this is a simulated dataset, we can verify the counterfactual value of Y^{cf} as the value of the individuality factor—here represented by E —is known.

Figure 6 shows the values of MAE/std for the 10 counterfactual queries created using 10 different values for X . Overall, note that the values of the measure are low for all cases, being always below 0.3. Looking at the results for both Y_1 and Y_2 in Fig. 6a, b for different values of α , note that as the individuality factor increases its multiplicative effect on y , the models learn better and have a lower counterfactual

error, converging to a value of ≈ 0.005 even when using extreme values of X for the counterfactual queries.

In Fig. 6c, d, when $\alpha = 0.0$, we have a different pattern: negative values of β show a higher error and variance among the curves, meaning that this type of function was harder for the algorithm to learn. The lowest values of error are obtained when β is in the interval ranging from 0.1 to 0.7, and here we also observe a lower variance between the values of the counterfactual query, but the differences of errors are still big when the counterfactual queries reach the extreme values.

Finally, Fig. 6e, f show the cases where the counterfactual queries reached their largest errors overall, getting near to 0.35 standard deviations. In these plots, the lowest values of MAE/std happened when $\alpha = \beta = 0.2$ for Y1 and Y2, i.e., when we had a situation close to linear function with an additive individuality factor. We can also note that although the error is a little higher when both α and β are 1—meaning a complete multiplicative and non-linear scenario—the highest error we obtained was lower than 0.25, which is small.

The results in Figs. 4 and 6 show that even in situations where the test error was high the model was still able to make good counterfactual predictions. For example, when $\alpha = 0.3$ and $\beta = 0.0$ we had a high test error in the factual prediction, with 0.58 standard deviations, but a low counterfactual query prediction with all standard deviation errors lower than 0.15. Also, when querying counterfactual for $\alpha = \beta = 0.7$, where we obtained a low test error and a counterfactual error above 0.2. These results show that there is not a necessary relation between the ability to generalize results and to answer good counterfactual questions correctly.

The results also show that when bigger changes are applied to the factual value, the counterfactual error tends to increase. This may be related to the standard problems in machine learning, such as model effectiveness or lack of representative training data. But it may also be related to the fact that the bottleneck can create non-continuous representations, which could lead to small changes in the factual values to cause very different impacts in the predictions. This is a well-known problem for auto-encoders, which share similarities with the proposed approach.

5.2 Mental health case study

This section applies the proposed framework to all three levels of Pearl's hierarchy of causal knowledge: (i) association, showing that, given the observed data we can predict a value close to the real one; (ii) intervention, measuring the impact of intervening in the whole population and seeing the statistic impact; (iii) individual counterfactual, by taking a sample of random people and seeing what would have happened if they had different preconditions. While in the first step we can quantitatively measure the correctness of the results, the second and third steps can only be evaluated qualitatively (Pearl 2009), as they are based in counterfactual outputs. Therefore, these results were validated by psychologists and psychiatrists.

The problem we deal with is related to two well-known instruments for measuring symptoms of mental health diseases and quality of life: the Brief Symptom Inventory (BSI) and WHOQOL-BREF (World Health Organization Instrument to evaluate

Table 2 Number of input and output variables and MAE measured for association relations with BSI and WHOQOL

BSI		WHOQOL		Train		Test	
#X ₁	y ₁	#X ₂	y ₂	y ₁	y ₂	y ₁	y ₂
8	<i>Somat.</i>	9	<i>Env</i>	0.0179	0.0079	0.0439	0.0081
8		12	<i>Phys</i>	0.0007	0.0007	0.0165	0.0217
8		10	<i>Psych</i>	0.0004	0.0005	0.0070	0.0094

Quality of Life: Brief Version), respectively. The scores related to these measures are generated based on a linear combination of answers provided by people in questionnaires.

BSI is an instrument designed to identify psychological symptoms in nine dimensions, including depression, anxiety, and somatization. There is a BSI score for each of these dimensions, accounting for the answers to specific questions. Here we focus on BSI-somatization. Somatization is the development and persistence of symptoms unexplained by medical causes, and it is well-known that somatization is the result of both depression and anxiety (Harding et al. 2015). The second score, WHOQOL-BREF, measures people's quality of life considering four domains: physical health, psychological, social relationships and environment. Its questions can also be mapped to the domains they are measuring. We work with all domains except social relationships, as it includes only three questions.

The counterfactual of these can only be done by finding each subject's individuality, since there are known latent variables that could not be measured, like genetics and life experiences (Friedman et al. 2014). It is known that the individuality affects both BSI and WHOQOL, which make these scores a perfect fit for the causal model proposed here.

There is plenty of evidence showing association between somatization and behavior symptoms (Wiborg et al. 2013; Hamdan et al. 2019). Besides evidence, there is a need to have causal models to understand more closely the alignment between the underlying biology and behavior, and how the change in scenario would impact on the outcome. Strategies such as the counterfactual might help in understanding the effects of these changes. It seems important not only to understand behavior but psychopathology.

The dataset was built from the Psychological Covid impact study, and involved a sample of 79 out of more than 100 questions answered online by 153,514 individuals from May-2020 to July-2020. It contains answers to all questions, which use the 5-points Likert scale from 0 (not at all) to 4 (extremely).

Because the scores we want to predict and generate counterfactuals are linear combinations of questions answered in the questionnaire, a regressor can easily predict the perfect score. This scenario is not appropriate to test the proposed framework, and hence we performed a data preparation step. First, as we have the mapping from questions to domains for both instruments, we calculate the correlation between the values of the calculated scores and the questions. We select the two questions most correlated to the score and disregard the others. These two variables are later merged with a set of other questions extracted from the causal graph, since we heavily rely on the correctness of the causality between the answers and scores.

Methodology The first step in our methodology is the inference of the causal graph using NOTEARS. For the BSI-Somatization score, the causal graph was inferred using all questions classified in the anxiety, depression and somatization categories. This choice was based on the causal graph produced by the authors in Grill et al. (2014). We used as input to our model the variables with a direct cause in the questions of somatization with weight greater than 0.05, where the weight corresponds to the strength of a causal relation. The 6 variables identified were merged with the two most correlated to the index, and the base model has an 8-variable input.

For WHOQOL we did not have any references to filter out variables, and hence we have used the answers to all questions to build the causal graph. For both the environmental WHOQOL (*Env*) and the Physical WHOQOL (*Phys*), the process was the same followed for BSI. The only difference was the cut-off value, which was 0.1 to *Phys*, as we observed the weight values were much higher for this graph. For WHOQOL Psychological (*Psych*), after building the causal graph we realized there was one variable that was caused by all others in the graph. Instead of adding all the cause variables to the graph, we opt to include a third variable that was directly associated with the score. The number of input variables used by each model is shown in Table 2, and their detailed description is given in Appendix A.1.

Association results Table 2 also shows the training and test MAEs for the base model using different combinations of BSI (y_1) and three variations of WHOQOL (y_2). Recall our model outputs two variables. The results show a test BSI MAE error between 0.007 and 0.0439, which is a low error considering that the values range from 0 to 1. QoL *Phys* obtained the highest test error among the QoL indices—0.0217—and *Env* the lowest. These results show the base model has the ability to learn associations and correctly fit the function for the factual prediction task.

Individual-level counterfactual results We now turn to the results for counterfactuals. In psychiatry, there is a constant pursue for understanding what actions would have created a better life style for a single individual. The experiments in this section shed light into that. We show that outcomes of the scores can be different given the same input, as different people will not react in the same manner to similar situations. This is why it is so important to have a deterministic individuality factor for each of the data samples, capable of generating a more personalized counterfactual query. While factual prediction methods learn a consistent output prediction for the same input, this is not necessarily true in counter-factual models. In this sense, an appropriate counterfactual should be capable of both predicting different outputs (considering the individuality as the latent variable) for the same input and having a different impact when counterfactually changing the values of input variables.

It is important to mention that, although both BSI and WHOQOL are influenced by the individuality of a person, studies investigating the association between BSI subscales and the WHOQOL domain scores generally find significant low to high negative correlations (Fellinger et al. 2005; Marques et al. 2017; Vertommen et al. 2018). One study found that BSI GSI (global score), depression, and anxiety showed the strongest correlations with psychological QOL, while BSI somatization was most

Table 3 Individual-level counterfactual analysis

ID	Score	Feature	F Value	F Score	CF Value	CF Score
1	BSI_Psy	BSI_45	0	0.0712	4	0.0713
2			0	0.0355	4	0.0712
1	Psychological	WHOQOL_19	3	0.6291	5	0.7355
2			3	0.7294	5	0.7838
3	Physical	WHOQOL_2	3	0.7399	5	0.7903
4			3	0.6716	5	0.6996
5	Environment	WHOQOL_5	4	0.7418	2	0.7218
6			4	0.8246	2	0.8044
7			4	0.7971	2	0.7785
8			4	0.7137	2	0.6921

strongly correlated with physical QOL, but correlations were weakened by controlling for factors such as participants' demographics (Vertommen et al. 2018). It is important to emphasize that psychological distress and quality of life are complex constructs building on human health and well-being, but are not the same (Evans et al. 2007; Friedman and Kern 2014). Note these studies look at correlation, and there is not a systematic work on causation. Hence, there is no evidence of a strong causal relation between BSI features and WHOQOL scores nor between WHOQOL features and BSI scores. For this reason, we only analyze the interventional impact between BSI features and BSI score and WHOQOL features and WHOQOL score. This lack of causal relation is in accordance with the proposed causal model in Fig. 2, where the outcomes do not have a causal edge between them, and the only expected information that they share is the individuality. Recall that a higher BSI score indicates a worse case of somatization, while a higher WHOQOL score indicates better health in the analyzed dimension.

Table 3 shows examples of predictions for factuais (F) and counterfactuals (CF) of users who have given the same answers to the questionnaire (i.e., have the same input variables) but to which the counterfactual network has predicted different values for BSI and WHOQOL scores. In the table, the IDs refer to individuals in the database, the score indicates the value that will be predicted by the counterfactual network, the feature is the answer from the questionnaire which has its value being analyzed, "F" stands for the factual value and "CF" to the counterfactual values.

Notice that individuals with IDs 1 and 2 in Table 3 have the same input values when the network predicts values for BSI Somatization and WHOQOL Psychological, but the method predicted their score for BSI as 0.0712 for ID 1 and 0.0355 for ID 2 and WHOQOL Psychological being 0.6333 and 0.7294, respectively. These values show that individual 2 has a better psychological quality of life and, while they both have a low somatization, the score for individual 1 is a little higher. Two counterfactual questions were asked: the first is what the BSI score would have been if they were in a situation where their BSI 45 ("spells of terror") answer would have been 4. The result shows that while the BSI of individual 2 increased to 0.0712 the BSI of individual 1 had almost no change, rising to 0.0713. When analyzing the WHOQOL we see that

the score of individual 2 had a small alteration (0.06) when asking what would have happened if (s)he was in a situation where hi(e)s “self satisfaction” (WHOQOL 19) would have been 5. For individual 1 the score increase in 0.10. This shows that not only the algorithm is capable of predicting different values from same input variables, as it is also capable of finding complex individuality impact in the counterfactual, where both BSI and WHOQOL change in different ways.

Regarding individuals with IDs 3 and 4, they have the same input values when the network predicts values for BSI Somatization and WHOQOL Physical, but the method predicted their score for WHOQOL Physical as 0.7399 and 0.6716 respectively. The data also shows that when predicting the counterfactual for a scenario where WHOQOL 2 (Health) would have value 5, individual 3 would have have a new Phys. WHOQOL of 0.7903, while individual 4 would have it as 0.6996.

The table also shows comparisons between individuals of ID 5, 6, 7, and 8 when predicting BSI Somatization and WHOQOL Environment. Note that the four of them had the same input and were predicted with different outputs for the two scores. This is an interesting case, because when predicting the counterfactual of WHOQOL 5 (“life enjoyment”) equals to 2, we can see that they all varied around their factual predicted value, but the variance was approximately the same—0.02 for all of them.

The results showed that the deterministic factual prediction can predict different outputs from the same factual inputs, and also provides different counterfactual impacts when applied similar changes. These results are expected: Friedman et al. (2014) exemplify this by using the serotonin levels in the nervous system, which have genetic basis but are also alterable by life circumstances. Serotonin levels also affect personality, i.e., neuroticism and conscientiousness and help regulate core bodily functions necessary for good health, i.e., appetite and sleep. The results in McCaffery et al. (2006) show that the covariation of depressive symptoms and coronary artery disease may be due to a common vulnerability, indicating that different people could have individuality’s that make them respond differently to similar situations.

The individual counterfactual analysis brings insights to professionals of what can be done to help a specific patient. The ability of understanding what could be changed for these individuals can develop strategies to improve treatment and anticipate actions. If the specialist knows an individual will have its BSI lowered the most by increasing a specific indicator, it could work on that indicator, e.g. sleep or social support.

Population counterfactual results In these experiments we look at the impact of actions in the whole population. In other words, after changing the values of the variables of interest for each individual, as previously reported, we calculate the percentage of people that had their BSI and WHOQOL scores increased, lowered, or maintained.

To analyze the impact of interventions in the whole population we tested extreme situations: the counterfactual query considering the value of the feature of interest being the minimum and the maximum value. People whose original responses in the questionnaire were equal to the extreme value analyzed were removed from this analysis as they would not reflect changes in the scores. We consider scores change if

	Min Value	Min Value	Max Value	Max Value
	▼	▲	▼	▲
Nervousness	8.37	0.1	0.01	17.41
Suicidal thoughts	13.53	0.08	0.01	73.98
Scared for no reason	13.38	0.06	0.02	37.13
Feeling tense	4.29	0.08	0.01	12.22
Spells of terror	31.37	0.01	0.01	82.68
Feeling restless	15.78	0.07	0.02	50.08

(a) Changes in BSI with Psych WHOQOL.

	Min Value	Min Value	Max Value	Max Value
	▼	▲	▼	▲
Nervousness	0.37	0.1	0.01	1.08
Suicidal thoughts	0.15	0.46	0.15	0.82
Scared for no reason	0.26	0.04	0.09	2.39
Feeling tense	0.09	0.28	0.02	0.51
Spells of terror	1.66	0.3	0.08	6.86
Feeling restless	0.19	0.13	0.07	1.52

(b) Changes in BSI with Phys WHOQOL.

	Min Value	Min Value	Max Value	Max Value
	▼	▲	▼	▲
Nervousness	32.96	21.6	19.82	33.83
Suicidal thoughts	45.67	34.44	13.42	62.0
Scared for no reason	43.69	28.96	18.26	57.25
Feeling tense	31.92	24.91	17.98	44.19
Spells of terror	55.62	27.6	17.65	49.69
Feeling restless	43.55	29.07	16.66	53.68

(c) Changes in BSI with Env WHOQOL.

	Min Value	Min Value	Max Value	Max Value
	▼	▲	▼	▲
Everyday energy	96.21	1.21	0.89	81.48
Leisure Opportunity	80.57	0.95	1.2	64.56
Sleep	88.88	4.44	5.04	76.91
Capacity for work	61.63	2.73	4.15	14.13
Personal relationships	88.19	6.34	4.7	72.22
Friends' support	90.94	4.19	2.9	68.32
Satisf. transport	64.77	5.77	3.63	11.93

(d) Changes in Psych WHOQOL with BSI.

Fig. 7 Heatmaps showing populational changes when counterfactually changing values of variables. First and third columns show the percentage of the population that has the value of the score (indicated in the caption) decreased when the variable in the line receives minimum and maximum values, respectively. Second and fourth columns show the same percentage of the population that have their score values increased with these changes

their values increase or decrease in at least 0.01 (1%).⁵ The results are summarized in Fig. 7 and a detailed version can be found in Table 6, Appendix 1. In the heatmaps in the figure, each line corresponds to a variable (a question in the questionnaire, also detailed described in Appendix A.2). They have the original value of their answers changed to the minimum (0) and maximum (4) value. The first and third columns of each map show the percentage of the population that have their score value (indicated in the caption) *decreased* when the variable in the line received the minimum and maximum values, respectively. The second and fourth columns show the same percentage of the population that have their score values *increased* with these same changes: when the variable in the line received the minimum and maximum values.

We first look at the results of BSI when counterfactual queries are executed considering the three different WHOQOL categories (first column in Fig. 7). Note that the impact of the results vary significantly from one model to the other. These variations are expected because, although the BSI score comes from the same function, each model captures different causal relations when combined with WHOQOL, which will lead the model to learn distinct individuality factors.

⁵ Other variations were also tested. Values over 5% include almost all individuals of the population, while lower values were more difficult to be justified since they could simply be capturing the model's prediction variation.

There is plenty of evidence showing association between somatization and behavior symptoms (Wiborg et al. 2013; Hamdan et al. 2019). A first observation that can be easily validated by a specialist is the impact of “suicidal thoughts” (BSI 9) in Fig. 7a, which increases the value of BSI for 78% of the population when it receives the higher possible score. Wiborg et al. (2013) showed that suicidality seems to be a substantial problem in primary care patients with somatoform disorders. They also argued that the perception of these dysfunctional illness may have a vital role in the understanding and management of active suicidal ideation. Hamdan et al. (2019) observed that the additional factors that bring the highest risk for suicide ideation for those who bereaved are symptomatology, post-traumatic stress disorder (PTSD), *somatization*, lower perceived social support, and secular religious orientation.

Similarly, in this same model the somatization score increased in more than 80% of the population when setting the max counterfactual value for “spells of terror” (BSI 45). This result can be seen as an interpretation of the work of Carmassi et al. (2020), which shows that physical symptoms are a common manifestation of potentially treatable psychiatric disorders. They also gathered results from other works that show that up to 50% of patients with somatic symptoms have sufficient criteria for a diagnosis of anxiety and mood disorder (DelloSo et al. 2009). This is in accordance with our results, where approximately 50% of the population was affected by a change to the maximum value in the answers for the question related to “feeling restless” (BSI 49) and approximately 37% was affected when increasing the value of “Scared for no reason” (BSI 12), which are symptoms of anxiety.

The results in Fig. 7b, in contrast, where BSI was combined with the physical WHOQOL, show little populational impact. This can be a reflex of the variables used to predict BSI, since we cannot guarantee the independence of these features and, most importantly, good predictors are not necessarily good causes. Conversely, when dealing with BSI and WHOQOL environment (Fig. 7c), we observe that all variables significantly impacted somatization, with populational changes varying from 32% to 62% when both reducing or increasing the values of scores of different variables. In these results, “suicidal thoughts” had again the largest impact in the population somatization.

Observe that, counter-intuitively, we have a significant increase/decrease in BSI values when changing variables to their min/max value (e.g., “suicidal thoughts” with the minimum value decreases somatization for 45% of the population (which is expected), but increases for 35%; similarly, decreasing the values for variables “nervousness”, “scared for no reason”, “feeling tense” and “spells of terror” had the BSI values increased for roughly 20% of the population. In this respect, there are some studies linking the changes in somatization to different expressions and suppression of anger feelings for men and women, respectively Liu et al. (2011). The study also showed the link between childhood trauma and adult somatic symptoms was mediated by insecure attachment. However, the mechanism which links this insecure attachment with somatization is still poorly understood. Koh et al. (2005) and Watson (1989) also showed that proneness to experiencing negative emotions and suppression of negative emotions are associated with somatoform disorders.

Turning now to the results of WHOQOL, we explore the results of WHOQOL psychological, shown in Fig. 7d. The results for the physical and environment WHOQOL

can be found in Appendix 1. For Psych WHOQOL, note that reducing the values of all variables had a big impact on quality of life for the whole population. These results are consistent with the ones reported in Killgore et al. (2020), where the authors found that average resilience was lower than published norms during the pandemics, but was greater among those who tended to get outside more often, exercise more, perceive more social support from family, friends and significant others, sleep better, and pray more often. The significant impact of sleep it is also a result that is found in the literature, where Jin et al. (2014) found that better sleep status was related to lower social stress, better management of stress and good social support. At the same time, when increasing the values of the variables to the maximum value, all variables increased their values for WHOQOL.

6 Conclusions and future work

This work introduced a new framework for counterfactual inference, where we model a multivariate regression problem that considers both observed variables and a latent variable φ , which represents the individuality factor. The individuality factor is inherent to a unit in our data, and needs to be learned.

The framework is capable of supporting the three levels of the SCM hierarchy of causal knowledge, being able to make associations, interventions and counterfactual. The framework has three main components, being our main contribution the ICoNet, a network that is trained to learn the individuality factor and can later be used to answer counterfactual queries.

Results in synthetic data show the network learns the individuality factor with very small error, and effectively uses the learned individuality to make counterfactuals. When using the mental health dataset, the counterfactual found by the model were all validated by specialists, which analyzed both intuitive as well as counter-intuitive findings, and corroborated the usefulness of the approach for clinical interventions on patient level.

The proposed framework can be applied to other health scenarios where the disease is influenced by individual characteristics, including different types of cancer and treatment, or how a person's anatomy changes if they have a different trait (e.g., as done by Pawlowski et al. (2020) for the brain). But health is not the only scenario where counterfactual queries may be useful.

In any application domain where the answer to *what-if* questions can help interventions and change behaviors or other unmeasurable variables are strong candidates. Having available data we can, for example, estimate the performance of students in online courses based on their grades and system interactions, and plan interventions to improve overall performance. In robotics, we can improve the robustness of a robot control system by training it using counterfactuals (Smith and Ramamoorthy 2020), to name a few other potential application scenarios.

Finally, while we have worked so far with simple counterfactual queries, which focused in changing a single variable, in the future the use of *recourse* (Karimi et al. 2021) can allow us to verify combinations of variables changes. Another possible improvement is to create a continuous bottleneck for the individuality extraction,

as we cannot currently guarantee that small individuality changes will not have big impacts in the model.

Acknowledgements This work was partially funded by the Brazilian Research Agencies CNPq, FAPEMIG, and CAPES, and by the projects Masweb, INCT-Cyber and CIIA-Saúde (Innovation Center on Artificial Intelligence for Health).

A Appendix

A.1 WHOQOL and BSI elements

Tables 4 and 5 present the questions considered to describe the individuals WHOQOL (WH Organization 1996) and BSI (Derogatis 1993), respectively.

Additional results

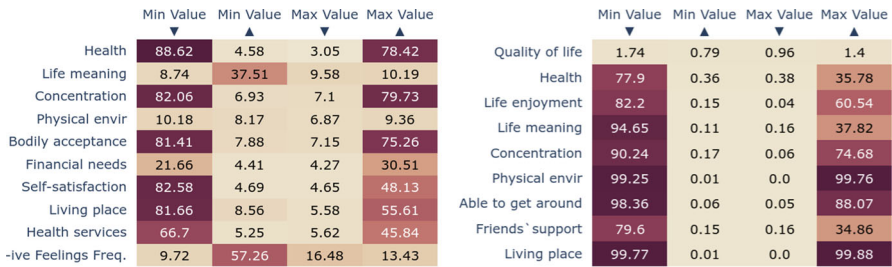
Population counterfactual results Apart from the results of psychological WHOQOL, discussed in Sect. 5, we have also analyzed the results of physical and environment

Table 4 WHOQOL questions considered by the model. All questions are answered following a range from 1 to 5, and scores in positive direction, i.e., higher scores denote higher quality of life

ID	WHOQOL question
1	How would you rate your quality of life?
2	How satisfied are you with your health?
5	How much do you enjoy life?
6	To what extent do you feel your life to be meaningful?
7	How well are you able to concentrate?
9	How healthy is your physical environment?
10	Do you have enough energy for everyday life?
11	Are you able to accept your bodily appearance?
12	Have you enough money to meet your needs?
14	To what extent do you have the opportunity for leisure activities?
15	How well are you able to get around?
16	How satisfied are you with your sleep?
17	How satisfied are you with your ability to perform your daily living activities?
18	How satisfied are you with your capacity for work?
19	How satisfied are you with yourself?
20	How satisfied are you with your personal relationships?
22	How satisfied are you with the support you get from your friends?
23	How satisfied are you with the conditions of your living place?
24	How satisfied are you with your access to health services?
25	How satisfied are you with your transport?
26	How often do you have negative feelings such as blue mood, despair, anxiety, depression?

Table 5 BSI topics considered by the model. All questions are answered following a 5-point scale of distress, ranging from “not-at-all” to “extremely”

ID	BSI factor
1	Nervousness or shakiness
9	Thoughts of ending your life
12	Suddenly scared for no reason
23	Nausea or upset stomach
37	Feeling weak in parts of your body
38	Feeling tense or keyed up
45	Spells of terror or panic
49	Feeling so restless you could not sit still



(a) Changes in Phys WHOQOL with BSI. **(b)** Changes in Env WHOQOL with BSI.

Fig. 8 Heatmaps showing populational changes when counterfactually changing values of variables. First and third columns show the percentage of the population that has the value of the score (indicated in the caption) decreased when the variable in the line receives minimum and maximum values, respectively. Second and fourth columns show the same percentage of the population that have their score values increased with these changes

WHOQOL. In contrast with the results of psychological WHOQOL, for physical WHOQOL (Fig. 8a) all results are intuitive and reported in the literature, except from “life meaning” (WHOQOL 6) and “Negative feelings Frequency” (WHOQOL 26), where reducing the sense of life meaning improves quality of life for 37% of the population. These findings go into the direction of the affirmations of Friedman and Kern (2014), who said that currently strong empirical support is lacking for sense of meaning as a vital factor in future health. However, the authors list it as a promising direction, especially because there is considerable evidence that persistent striving for meaningful accomplishment is indeed a key pathway to health and longevity.

For WHOQOL environment, “Health satisfaction” (WHOQOL 2), “having a meaningful life” (WHOQOL 6) and “Friend’s support” (WHOQOL 22) are known features related with well-being and longevity. The changes of these scores to minimum values impacted many people (see Fig. 8b). Friedman and Kern (2014) support this by showing results that link longevity with close relationships, having mastery over one’s environment, experiencing spirituality, engaging in life, and shows that maintaining positive relationships with others improve well-being and general health.

Table 6 shows the values of the counterfactuals when considering the whole population that were summarized in the heatmaps previously shown in this paper.

Table 6 Percentage of variation in the BSI and WHOQOL scores as counterfactual interventions are made in the variables indicated in the Id column

Id	Physical intervention			Psychological intervention			Environmental intervention			=										
	Min	Max	>	Min	Max	>	Min	Max	>											
<i>BSI</i>																				
1	0.1	0.37	99.53	1.08	0.01	98.91	1	0.1	8.37	91.53	17.41	0.01	82.58	1	21.6	32.96	45.44	33.83	19.82	46.35
9	0.46	0.15	99.39	0.82	0.15	99.03	9	0.08	13.53	86.39	73.98	0.01	26.01	9	34.44	45.67	19.9	62	13.42	24.57
12	0.04	0.26	99.7	2.39	0.09	97.51	12	0.06	13.38	86.56	37.13	0.02	62.85	12	28.96	43.69	27.35	57.25	18.26	24.49
*23	0	99.98	0.01	100	0	0	23	0.02	76.38	23.61	99.54	0.01	0.45	23	6.69	88.88	4.43	94.91	3.43	1.66
*37	0	99.99	0.01	100	0	0	37	0.04	80.44	19.51	99.86	0.01	0.13	37	6.11	89.9	3.99	96.5	2.35	1.15
38	0.28	0.09	99.63	0.51	0.02	99.47	38	0.08	4.29	95.63	12.22	0.01	87.78	38	24.91	31.92	43.16	44.19	17.98	37.82
45	0.3	1.66	98.04	6.86	0.08	93.06	45	0.01	31.37	68.62	82.68	0.01	17.31	45	27.6	55.62	16.78	49.69	17.65	32.67
49	0.13	0.19	99.68	1.52	0.07	98.41	49	0.07	15.78	84.15	50.08	0.02	49.9	49	29.07	43.55	27.39	53.68	16.66	29.66
<i>WHOQOL</i>																				
2	4.58	88.62	6.8	78.42	3.05	18.52	*6	0.31	99.2	0.5	93.14	0.1	6.76	1	0.79	1.74	97.47	1.4	0.96	97.64
6	37.51	8.74	53.75	10.19	9.58	80.23	10	1.21	96.21	2.58	81.48	0.89	17.63	2	0.36	77.9	21.74	35.78	0.38	63.84
7	6.93	82.06	11.01	79.73	7.1	13.16	14	0.95	80.57	18.47	64.56	1.2	34.24	5	0.15	82.2	17.65	60.54	0.04	39.42
9	8.17	10.18	81.65	9.36	6.87	83.78	16	4.44	88.88	6.68	76.91	5.04	18.06	6	0.11	94.65	5.24	37.82	0.16	62.02
*10	1.4	96.64	1.96	98.98	0.4	0.62	18	2.73	61.63	35.64	14.13	4.15	81.72	7	0.17	90.24	9.59	74.68	0.06	25.26
11	7.88	81.41	10.71	75.26	7.15	17.58	*19	0.13	99.54	0.33	99.66	0.08	0.26	*9	0.01	99.25	0.74	99.76	0	0.23
12	4.41	21.66	73.93	30.51	4.27	65.22	20	6.34	88.19	5.46	72.22	4.7	23.08	15	0.06	98.36	1.58	88.07	0.05	11.88
*17	0.65	98.43	0.92	99.63	0.15	0.21	22	4.19	90.94	4.87	68.32	2.9	28.79	22	0.15	79.6	20.25	34.86	0.16	64.98
19	4.69	82.58	12.73	48.13	4.65	47.22	25	5.77	64.77	29.46	11.93	3.63	84.43	*23	0.01	99.77	0.23	99.88	0	0.12
23	8.56	81.66	9.78	55.61	5.58	38.81	*26	0.22	99.28	0.5	99.59	0.13	0.27							
24	5.25	66.7	28.05	45.84	5.62	48.54														
26	57.26	9.72	33.02	13.43	16.48	70.09														

Ids proceed by an (*) indicate the variables directly used to generate the score

References

- Belkin M, Hsu D, Ma S, Mandal S (2019) Reconciling modern machine-learning practice and the classical bias-variance trade-off. *PNAS* 116(32):15849–15854
- Bica I, Alaa A, Van Der Schaar M (2020) Time series deconfounder: estimating treatment effects over time in the presence of hidden confounders. In: *ICML*, pp 884–895
- Buesing L, Weber T, Zwols Y, Heess N, Racaniere S, Guez A, Lespiau JB (2018) Woulda, coulda, shoulda: counterfactually-guided policy search. In: *ICLR*
- Bzdok D, Meyer-Lindenberg A (2018) Machine learning for precision psychiatry: opportunities and challenges. *Biol Psychiatry Cogn Neurosci Neuroimaging* 3(3):223–230
- Carmassi C, DellOste V, Cordone A, Pedrinelli V et al (2020) Relationships between somatic symptoms and panic-agoraphobic spectrum among frequent attenders of the general practice in Italy. *J Nerv Ment Dis* 208(7):540–548
- Chickering DM (1996) Learning Bayesian networks is np-complete. In: *Learning from data*. Springer, pp 121–130
- Chickering M, Heckerman D, Meek C (2004) Large-sample learning of Bayesian networks is np-hard. *J Mach Learn Res* 5:1287–1330
- DellOso L, Bazzichi L, Consoli G, Carmassi C, Carlini M, Massimetti E, Giacomelli C, Bombardieri S, Ciapparelli A (2009) Manic spectrum symptoms are correlated to the severity of pain and the health-related quality of life in patients with fibromyalgia. *Clin Exp Rheumatol* 27(5):S57
- Derogatis LR (1993) *BSI brief symptom inventory*. Administration, scoring, and procedures manual, 4th edn. National Computer Systems, Minneapolis, MN
- Eberhardt F (2017) Introduction to the foundations of causal discovery. *Int J Data Sci Anal* 3(2):81–91
- Evans S, Banerjee S, Leese M, Huxley P (2007) The impact of mental illness on quality of life: a comparison of severe mental illness, common mental disorder and healthy population samples. *Qual Life Res* 16(1):17–29
- Fellinger J, Holzinger D, Dobner U, Gerich J, Lehner R, Lenz G, Goldberg D (2005) Mental distress and quality of life in a deaf population. *Soc Psychiatry Psychiatr Epidemiol* 40(9):737–742
- Friedman HS, Kern ML (2014) Personality, well-being, and health. *Annu Rev Psychol* 65:719–42
- Friedman HS, Kern ML, Hampson SE, Duckworth AL (2014) A new life-span approach to conscientiousness and health: combining the pieces of the causal puzzle. *Dev Psychol* 50(5):1377
- Graham L, Lee CM, Perov Y (2019) Copy, paste, infer: a robust analysis of twin networks for counterfactual inference. In: *NeurIPS19 CausalML workshop*
- Grill E, Schäffler F, Huppert D, Müller M, Kapfhammer HP, Brandt T (2014) Self-efficacy beliefs are associated with visual height intolerance: a cross-sectional survey. *PLoS ONE* 9(12):e116220
- Guidotti R, Monreale A, Ruggieri S, Pedreschi D, Turini F, Giannotti F (2018) Local rule-based explanations of black box decision systems. [arXiv:1805.10820](https://arxiv.org/abs/1805.10820)
- Hamdan S, Berkman N, Lavi N, Levy S, Brent D (2019) The effect of sudden death bereavement on the risk for suicide: the role of suicide bereavement. *Crisis J Crisis Interv Suicide Prev* 41:214–224
- Harding KA, Murphy KM, Mezulis A (2015) Cognitive mechanisms reciprocally transmit vulnerability between depressive and somatic symptoms. *Depress Res Treat*. <https://doi.org/10.1155/2015/250594>
- Jin Y, Ding Z, Fei Y, Jin W, Liu H, Chen Z, Zheng S, Wang L, Wang Z, Zhang S et al (2014) Social relationships play a role in sleep status in Chinese undergraduate students. *Psychiatry Res* 220(1–2):631–638
- Karimi AH, Schölkopf B, Valera I (2021) Algorithmic recourse: from counterfactual explanations to interventions. In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pp 353–362
- Killgore WD, Taylor EC, Cloonan SA, Dailey NS (2020) Psychological resilience during the covid-19 lockdown. *Psychiatry Res* 291:113216
- Koh KB, Kim DK, Kim SY, Park JK (2005) The relation between anger expression, depression, and somatic symptoms in depressive disorders and somatoform disorders. *J Clin Psychiatry* 66(4):485–491
- Le T, Wang S, Lee D (2020) Grace: generating concise and informative contrastive sample to explain neural network models prediction. In: *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery and data mining*. Association for Computing Machinery, New York, NY, USA, KDD 20, pp 238–248. <https://doi.org/10.1145/3394486.3403066>
- Liu L, Cohen S, Schulz MS, Waldinger RJ (2011) Sources of somatization: exploring the roles of insecurity in relationships and styles of anger experience and expression. *Soc Sci Med* 73(9):1436–1443

- Locatello F, Bauer S, Lucic M, Raetsch G, Gelly S, Schölkopf B, Bachem O (2019) Challenging common assumptions in the unsupervised learning of disentangled representations. In: ICML, pp 4114–4124
- Louizos C, Shalit U, Mooij J, Sontag D, Zemel R, Welling M (2017) Causal effect inference with deep latent-variable models. In: NeurIPS
- Marques DR, Meia-Via AMS, da Silva CF, Gomes AA (2017) Associations between sleep quality and domains of quality of life in a non-clinical sample: results from higher education students. *Sleep Health* 3(5):348–356
- McCaffery JM, Frasure-Smith N, Dubé MP, Thérroux P, Rouleau GA, Duan Q, Lespérance F (2006) Common genetic vulnerability to depressive symptoms and coronary artery disease: a review and development of candidate genes related to inflammation and serotonin. *Psychosom Med* 68(2):187–200
- Miao W, Geng Z, Tchetgen Tchetgen EJ (2018) Identifying causal effects with proxy variables of an unmeasured confounder. *Biometrika* 105(4):987–993. <https://doi.org/10.1093/biomet/asy038>
- Miller T (2020) Contrastive explanation: a structural-model approach. [arXiv:1811.03163](https://arxiv.org/abs/1811.03163)
- Montgomery MR, Gragnoli M, Burke KA, Paredes E (2000) Measuring living standards with proxy variables. *Demography* 37(2):155–174. <https://doi.org/10.2307/2648118>
- Nahum-Shani I, Smith SN, Spring BJ, Collins LM, Witkiewitz K, Tewari A, Murphy SA (2018) Just-in-time adaptive interventions (JITAs) in mobile health: key components and design principles for ongoing health behavior support. *Ann Behav Med* 52(6):446–462
- Paredes P, Gilad-Bachrach R, Czerwinski M, Roseway A, Rowan K, Hernandez J (2014) Poptherapy: coping with stress through pop-culture. In: Proceedings of the 8th international conference on pervasive computing technologies for healthcare, pp 109–117
- Pawlowski N, Coelho de Castro D, Glocker B (2020) Deep structural causal models for tractable counterfactual inference. In: NeurIPS
- Pearl J (2009) *Causality*. Cambridge University Press, Cambridge
- Pearl J, Glymour M, Jewell NP (2016) *Causal inference in statistics: a primer*. Wiley, New York
- Schölkopf B (2019) Causality for machine learning. [arXiv preprint arXiv:1911.10500](https://arxiv.org/abs/1911.10500)
- Shalit U, Johansson FD, Sontag D (2017) Estimating individual treatment effect: generalization bounds and algorithms. In: ICML, pp 3076–3085
- Shen X, Ma S, Vemuri P, Simon G (2020) Challenges and opportunities with causal discovery algorithms: application to Alzheimers pathophysiology. *Sci Rep* 10(1):1–12
- Singla S, Pollack B, Chen J, Batmanghelich K (2020) Explanation by progressive exaggeration. In: ICLR
- Smith SC, Ramamoorthy S (2020) Counterfactual explanation and causal inference in service of robustness in robot control. In: 2020 Joint IEEE 10th international conference on development and learning and epigenetic robotics (ICDL-EpiRob). IEEE, pp 1–8
- Stegle O, Janzing D, Zhang K, Mooij JM, Schölkopf B (2010) Probabilistic latent variable models for distinguishing between cause and effect. *NeurIPS* 23:1687–1695
- Tennant PW, Harrison WJ, Murray EJ, Arnold KF, Berrie L, Fox MP, Gadd SC, Keeble C, Ranker LR, Textor J, et al. (2019) Use of directed acyclic graphs (DAGs) in applied health research: review and recommendations. [medRxiv](https://arxiv.org/abs/1908.08111)
- Vallet F, Cailton JG, Refregier P (1989) Linear and nonlinear extension of the pseudo-inverse solution for learning Boolean functions. *EPL (Europhys Lett)* 9(4):315
- Vertommen T, Kampen J, Schipper-van Veldhoven N, Uzieblo K, Van Den Eede F (2018) Severe interpersonal violence against children in sport: associated mental health problems and quality of life in adulthood. *Child Abuse Negl* 763:459–468
- Wachter S, Mittelstadt B, Russell C (2018) Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harv J Law Technol* 31:841–887. <https://doi.org/10.2139/ssrn.3063289>
- Watson D (1989) Strangers ratings of the five robust personality factors: evidence of a surprising convergence with self-report. *J Pers Soc Psychol* 57(1):120
- World Health Organization, et al. (1996) WHOQOL-BREF: introduction, administration, scoring and generic version of the assessment: field trial version, December 1996. Technical report, World Health Organization
- Wiborg JF, Gieseler D, Fabisch AB, Voigt K, Lautenbach A, Löwe B (2013) Suicidality in primary care patients with somatoform disorders. *Psychosom Med* 75(9):800–806
- Wooldridge JM (2009) On estimating firm-level production functions using proxy variables to control for unobservables. *Econ Lett* 104(3):112–114. <https://doi.org/10.1016/j.econlet.2009.04.026>

- Zhang J, Bareinboim E (2020) Designing optimal dynamic treatment regimes: A causal reinforcement learning approach. In: ICML, pp 11012–11022
- Zhang C, Bengio S, Hardt M, Recht B, Vinyals O (2016) Understanding deep learning requires rethinking generalization. arXiv preprint [arXiv:161103530](https://arxiv.org/abs/1611.03530)
- Zhang Y, Bellot A, Schaar M (2020) Learning overlapping representations for the estimation of individualized treatment effects. In: International conference on artificial intelligence and statistics, PMLR, pp 1005–1014
- Zheng X, Aragam B, Ravikumar PK, Xing EP (2018) DAGs with no tears: continuous optimization for structure learning. *NeurIPS* 31:9472–9483

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.