Correct deconfounding enables causal machine learning for precision medicine and beyond

Vera Komeyer[1,2,3], Prof. Dr. Simon B. Eickhoff[1,2], Dr. Charles Rathkopf[1], Prof. Dr. Christian Grefkes[4,5,6], Dr. Kaustubh R. Patil[1,2], Dr. Federico Raimondo[1,2*]*

[1]Institute of Neuroscience and Medicine, Brain and Behaviour (INM-7), Research Centre Juelich, Juelich, Germany
[2]Institute of Systems Neuroscience, Medical Faculty, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[3]Department of Biology, Faculty of Mathematics and Natural Sciences, Heinrich Heine University Duesseldorf, Duesseldorf, Germany
[4]Goethe University Frankfurt and University Hospital Frankfurt, Department of Neurology, Frankfurt (Main), Germany
[5]Department of Neurology, University Hospital Cologne and Medical Faculty, University of Cologne, Cologne, Germany
[6]Institute of Neuroscience and Medicine, Cognitive Neuroscience (INM-3), Research Centre Juelich, Juelich, Germany
* Correspondence to Vera Komeyer (v.komeyer@fz-juelich.de) and Federico Raimondo (f.raimondo@fz-juelich.de)

Abstract

Artificial intelligence holds promise for individualized medicine. Yet, predictive models in the neurobiomedical domain suffer from a lack of generalizability and replicability so that transitioning models from prototyping to clinical applications still poses challenges. Key contributors to these challenges are confounding effects; in particular the oftentimes purely statistical perspective on confounding. However, complementing these statistical considerations with causal reasoning from domain knowledge can make predictive models a tool for causal biomedical inference beyond associative insights. Such causal insights give answers to biomedical questions of *how* and *why*, arguably what most biomedical investigations ultimately seek for. Here, we suggest a 5-step approach for targeted, context-informed deconfounding. We exemplify the 5-step approach with a real-world neurobiomedical predictive task using data from the UK Biobank. The core of this approach constitutes a bottom-up causal analysis to identify a correct set of deconfounders and the appropriate deconfounding method for a given causal predictive endeavour. Using the 5-step approach to combine causal with statistical confounder considerations can make predictive models based on observational (big) data a technique comparable to Randomized Control Trials (RCTs). Through causally motivated deconfounding we aim at facilitating the development of reliable and trustworthy AI as a medical tool. In addition, we aim to foster the relevance of low performing or even null result models if they originate from a "skilful interrogation of nature", i.e. a deconfounding strategy derived from an adequate causal and statistical analysis. Ultimately, causal predictive modelling through appropriate deconfounding can contribute to mutual recursive feedback loops of causal insights across disciplines, scales and species that enable the field to disentangle the cause-effect structure of neurobiomedical mechanisms.

1. Main

Machine Learning (ML) holds promise for personalized medicine and is increasingly employed in biomedical research and applications. ML workflows use large, high-dimensional and oftentimes observational data to arrive at predictive models to identify biomarkers of health and disease or to aid in diagnosis, prognosis and treatment choice, targeted to individuals[1–3]. Predictive modelling is thereby a prominent strategy to derive both, scientific insights regarding biomedical mechanisms as well as a clinical tool for precision-medicine.

Although promising, biomedical AI suffers from unreliable predictions[4–7], a lack of reproducibility and replicability, non-interpretability[8], and limited generalizability[9] of models. A key contributor to these challenges are confounding effects[10–12]. Classical examples of confounders include measurement artifacts[13–16], site effects[17], demographics[18–20], or lifestyle factors[21]. Large data, as required for AI applications, tend to be observational in nature. However, in observational data confounders must be accounted for by post-hoc statistical approaches, such as (linear) confounder regression[10,11,13,22–26]. In many biomedical disciplines it is common to correct for a conventionally established set of confounders, such as *sex* and *age*[5,10,27,28], without any justification[29–31]. If a justification is given, this is often in the form of a statistical association between the predictors and the confounder[29,32,33]. Reporting statistical associations appears appropriate when following the ubiquitous (but faulty – see **Box 3**) definition of confounders as any variable that correlates with the feature (predictor) and the target (outcome), but which's variance is of no interest[34,35]. Despite a variety of statistical methods for post-hoc confounder treatment, confounding still leads to - or is at least part of - the AI-challenges mentioned above. The reason being, treating confounding based on the above definition as a purely statistical notion, leads to confounding being dealt with purely by statistical means. However, confounding is not only a statistical notion, but also necessitates causal reasoning[36], on which we will elaborate within this paper.

Complementing statistical confounder considerations with causal reasoning from domain knowledge can make predictive models a tool for causal biomedical inference, going beyond associative insights. Here, we explain and exemplify how targeted, context-informed deconfounding in observational (big) data can make predictive models a technique comparable to Randomized Control Trials (RCTs). First, we distinguish between *high performance* and *understanding biology* models and highlight the role of confounding in this distinction. Second, by means of an exemplary predictive task we illustrate why and how ignoring causal reasoning while solely relying on correlative reasoning can lead to biased models and well-known paradoxes such as the Simpson's Paradox. As a solution, we discuss theoretically how to arrive at *understanding biology* models with (big) observational data through causal reasoning. We use the introduced real-world predictive example to illustrate an actionable 5-step approach to arrive at provisional causal models through targeted and informed deconfounding. Eventually, we discuss that it is particularly hard in the field of biomedical research to define a set of satisfying causal assumptions because of the inherently multi-dimensionality of biomedical mechanisms and close with suggestions on how to treat this dilemma.

2. The necessity for causal reasoning in predictive models and the role of confounders

2.1. Biomedical questions ultimately ask about the "why" and "how" of a phenomenon of interest

In the development of AI-tools, the medical usefulness and clinical trustworthiness of ML models is oftentimes (solely) judged based on a model's performance – "the higher the accuracy, the better the model", leading to a performance race in model development. Problematically, the achieved high performances oftentimes cannot be replicated under changing conditions. This makes previously high performing models fail in clinical deployment, i.e. models fail to generalize. The statistical solution is to avoid data distribution[37] and covariate shifts[38], by attempting to keep or make distributions of variables the same in the training and testing data. However, in real-world (medical) use-cases this cannot always be guaranteed. There is the demand for transportable and adaptable models between settings, i.e. under changing distributions. For example, a useful model should work in different hospitals, not just in the one on which's data it was trained. The demand is for models that can be used in the same way as for instance a glucose test, which gives the same results no matter where it is applied. Independent of the setting, it informs about glucose-tolerance and thereby supports diagnosis of diabetes. This is different for predictive models. Under new conditions, models must be trained again to learn a new prediction function as fitting a function to data is ultimately what any type of learning

technique can achieve[36]. While the *high performance* model operates based on learned patterns in data, the glucose test operates based on the knowledge of the underlying biological mechanism: Sustained higher blood glucose level after glucose intake can be caused by insulin resistance, i.e. diabetes type 2. In other words, the glucose test works under changing conditions because it is based on knowing the "why" and "how" of the biochemical mechanisms underlying diabetes, or put differently, knowing the underlying cause-effect relationship. Consequently, what is ultimately desired in biomedical predictive modeling are *understanding biology* models, that both, incorporate and enhance knowledge on causal biomedical effects. High performance aims should build on such causal models, because high performance based on valid biomedical mechanisms fosters model generalizability across different settings, which can improve trust in the usage of predictive models as biomedical tools.

Beyond biomedical models as clinical tools, arguably, biomedical questions often ultimately implicitly - even if not formulated explicitly - ask about the *why* and *how* of a phenomenon of interest. Why does person A have a higher hand grip strength (HGS) than person B? Why does person A suffer from depression, but a seemingly matching person B does not and why is the treatment in patient A successful but not in patient B? The problem though is, no matter how big, data are inherently "dumb about questions of why"[36]. The reason is that most predictive data-driven models are associative and observational in nature. However, asking *why* is a causal question that seeks for understanding of cause-effect relationships between variables. Given the well-known fact that correlation is not causation, it becomes clear that one cannot derive medical cause-effect relationships from a correlative (associative), observational approach, such as purely data-driven modeling. However, some correlations do imply causation. To disentangle if an association between a feature (predictor) and a target (outcome) does indeed imply causation, one needs to combine qualitative, causal information with quantitative, data information[36]. This does not only apply to causal predictive modeling but also domains such as structural equation modeling, i.e. some hypothesized causal structure has to be added to pure quantitative, associative interrogations[39]. Integrating causal assumptions can push ML techniques to allow for answering real-world *why* questions beyond quantifying associative patterns in data.

Box 1 – The ladder of causation

Causation can be distinguished into three levels of increasing causal insight: Seeing, doing and imagining[36] (Table B1).

Table B1. Ladder of causation (adapted from[36]).

Ladder Rung	Action	Learning Type	Questions	Examples	Gained Insight
Rung 3	Imagining 💭	**Counterfactuals**, imaging worlds that do not exist	What if X had not occurred, would Y have happened?	What if I had not smoked for the last 2 years, would I still have gotten lung cancer?	Understanding
Rung 2	Doing 🔨	**Intervention**, act by planning and learn from interventions	What would Y be if I do X?	If I take aspirin, will my headache be cured?	Causal mechanism
Rung 1	Seeing 👁️	Learning from **association**	How are two variables related?	What does a symptom say about a disease?	Correlative, pattern in the data

Rung one refers to the *seen world*. By investigating how two variables are related, new insights are derived from associations. This resembles the investigation of correlative patterns in data. Rung one analyses can answer observational, associative questions, for example in the 1700s being a sailor was associated with a higher risk of developing scurvy.

Rung two requires *doing* or *intervening*. Here, the interest lies in gaining more detailed understanding in cause-effect relationships by changing (intervening on) a predictor (feature) variable X and learning how an outcome (target) variable Y would change as an effect of manipulating X. Interventions allow to get insights into causal mechanisms. In a medical setting RCTs are the established means for interventional investigations. For example intervening on one ship with sailors having citrus fruits while another ship does not have could show that sailors on the citrus-fruit-ship didn't develop scurvy.

Rung three deals with a world that cannot be seen because it contradicts what is seen. Deriving insight on rung three requires imagining situations that do not exist, for example "would the outcome (target) Y have happened, if the predictor (feature) X had not occurred"? Rung three allows for not only seeing effects of interventions but understanding cause-effect mechanisms. For example, would a sailor S on the non-citrus-fruit-ship also not have developed scurvy if they (he) had had citrus fruits?

2.2. Not acknowledging the causal nature of confounders leads to paradoxes

Interventions allow to gain insights into causal mechanism beyond associative patterns in data and are therefore the next step towards answering questions of *why* (**Box 1**). Randomized Control Trials (RCTs) in medical experimentation are an established method to gain interventional insights because they implicitly take a causal note (**Box 2**). In contrast, confounder regression in predictive models stays correlative. However, not acknowledging the causal nature of confounders and other types of 3rd variables (**Box 3**) leads to a set of paradoxes such as Simpson's paradox (confounder bias) or Berkson's paradox (collider bias) (**Box 3**). As an example consider the supervised prediction of hand grip strength (HGS) from T1w-MRI derived grey matter volume (GMV) features in a large observational dataset such as the UK Biobank[40] (for methods see supplementary materials). A *vanilla* model without confounder considerations decently predicts HGS from parcellated GMV (**Fig. 1**a top). Following common practice of confounder removal, a second model can be built with linearly regressing out *sex* as a conventionally established confounder (also referred to as 3rd variable) (**Fig. 1**a bottom). Following the above given ubiquitous definition of a confounder as a variable that correlates with both, the features and the target, this decision could be backed up by the given point biserial correlation (statistical association) between *sex* and both HGS (r=.73) and GMV (r=.45) (**Fig. 1**b). The *vanilla* and the *sex*-adjusted model differ notably in their predictive performance ($R^2_{vanilla}$=.40 vs. R^2_{sex}=.03) (**Fig. 1**a). This high difference suggests that the good predictive performance of the *vanilla* model originated from a feature-target correlation that only exists without confounder regression. Such correlations – sometimes referred to as spurious correlations - can arise when two heterogenous populations are aggregated into one[36], known as Simpson's paradox. It occurs inter alia when the statistical result of the subgroups differs from the whole (aggregated) population (**Fig. 1**c) (see **Box 3** for definition). For example a drug happens to be bad for men and bad for women but good for *people*. In our scenario, a correlation of r=.44 between (unparcellated, whole brain) GMV and HGS in the aggregated population (male and female) in contrast to r_m=.22 and r_f=.15 in the two groups (**Fig. 1**b, d) suggests that the aggregation of the subpopulations creates a spurious correlation which the *vanilla* model leverages. In other words, by inappropriately combining two distinct populations (here: male and female), we created a supposedly good performing *vanilla* model whose success however was built on sex information, i.e. on a spurious correlation between the features and the target.

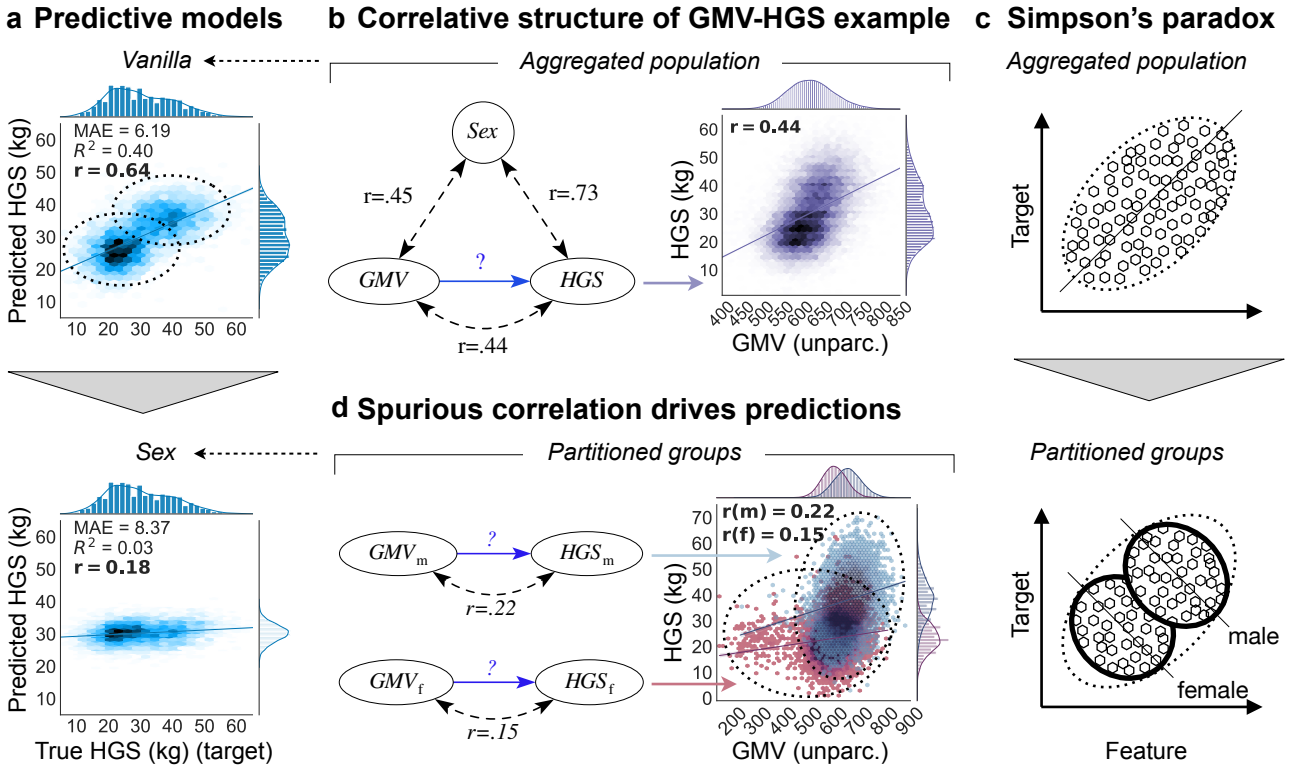

Fig. 1. Not acknowledging the causal nature of confounding can lead to paradoxes and spurious correlations that drive predictions. a) Supervised prediction of hand grip strength (HGS) from grey matter volume (GMV) with no confounder regression (*vanilla*) and regression of sex as confounder. b) Association of GMV and HGS in the aggregated population (male and female) with each other and with sex. c) Visualization of the Simpson's paradox. d) Association of GMV and HGS separately for males and females.

In the above example, partitioning the data seemed to be the right decision. However, aggregating the data (not adjusting for a 3rd variable) is not always wrong or partitioning the data (adjusting for a 3rd variable) is not always right. Rather, the right decision depends on the process that generated the data. This process needs to be understood individually for each predictive modeling task and this understanding necessitates integrating causal structures between variables. The data generating process cannot be revealed by correlative considerations alone because the correlative nature of a 3rd variable with the feature(s) and the target stays the same irrespective of the 3rd variable being a confounder, a mediator or a collider, but directionalities (causalities) differ (**Box 3**). For example, only given the correlation between GMV, *sex* and HGS (**Fig. 1**b) directionalities (i.e. causalities) cannot be distinguished. Consequently, it remains unclear whether *sex* is a confounder, a mediator or a collider. The occurrence of the Simpson's paradox when conditioning on (regressing out) *sex* suggested that here *sex* is a confounder because this paradox would not have occurred if *sex* were a mediator or collider (**Box 3**). Knowing that *sex* is a confounder, one needs to condition on *sex* to get insights into the causal path GMV → HGS. In contrast, conditioning on or regressing out a mediator would disable the causal path of interest (**Box 3**). Conditioning on a collider would even introduce a spurious correlation (Berkson's paradox, **Box 3**), for instance detectable through an increased accuracy. This means that conditioning on all statistically associated 3rd variables - maybe with a "better safe than sorry"-mindset - can lead to wrong insights because the correct decision for conditioning on a variable depends on the causal story not on the data. The Simpson's paradox alerts to cases where at least one of the statistical results - either from the aggregated data, the partitioned data, or both - cannot represent the causal effects. In the GMV-HGS example, the aggregated data does allow to investigate if GMV causes HGS.

Box 2 – Concepts and Terminologies for Causal Investigations

Investigations on causal inference require formal representation of causal concepts and assumptions. Causal diagrams or directed acyclic graphs (DAG), are used to express the known causal assumptions ("what we know"). Symbolic language supplements these diagrams by expressing the causal relationship to be found ("what we want to know")[36].

DAG: directed acyclic graph – "What we know"

A DAG is a circle-arrow picture. The circles represent variables, and the arrows represent directions of known or suspected causal relationships between two variables. X → Y in a DAG would mean that X is a direct cause of Y, i.e. the arrow implicitly says that some probability rule or function specifies how Y would change if X were to change or simplified, "Y listens to X". The rule according to which this change happens might either be known (e.g. previous research) or has to be estimated from data. However, often the structure of the DAG itself already enables to estimate causal relationships (simple or complicated, deterministic or probabilistic, linear or nonlinear). For example, a barometer reading B tracks the atmospheric pressure P. We know that it is the pressure P that causes the barometer reading to change, i.e. P → B, and not the other way around. The mere formular P=B/k wouldn't have revealed this causal directionality. Hence, a DAG depicts qualitatively the cause-effect forces that operate in the environment and that shape the data generated[36].

Formal probabilistic language and the do-operator - "What we want to know"

Types of probabilities

1. **P(S=s)**: The probability of the Variable S taking the value s. E.g. the probability of people in a café ordering scones.
2. **P(T=t, S=s)**: The probability of simultaneously T taking the value t and S taking the value s. E.g. the probability of people in a café ordering scones and tea.
3. **P(T=t | S=s)**: The probability that T=t conditional on finding S=s, i.e. the population distribution of T among individuals whose S value is s. E.g. the probability of people who have ordered scones to also order tea. ⇒ a distribution based on an observation.
4. **P(T=t | do(S=s)):** The probability that T=t when we intervene to make S=s, i.e. the population distribution of T if everyone in the population had their S value fixed at s. E.g. the probability to order tea when the person was "forced" to order scones. ⇒ do(s) creates a distribution by an intervention.

Medical example using probabilistic language to express a causal question (query)

Question: What is the effect of a drug (D) on lifespan (L)?

Formal expression: P(L | do(D))

In words: What is the probability (P) that a typical patient would survive L years if made to take the drug?

The do-operator formalizes interventional (treatment) questions and hence corresponds to what is measured in clinical trials. The "control" patients in the above example would be described as P(L | do(not-D)). It is important to note that P(L | D) may be different from P(L | do(D)). P(L | D) notes the observed probability of Lifespan L among patients who voluntarily take the drug (D) (standard conditional probability), while P(L | do(D)) is the probability of Lifespan L of patients made to take the drug. It is hence the fundamental difference between seeing and doing. In the barometer example from above, seeing the barometer reading (B) to fall increases the probability of a storm (lower atmospheric pressure) (P(P | B)). However, forcing the barometer read to fall does not affect the probability of a storm (P(P | do(B)). This means that P(P | B) ≠ P(P | do(B)).

Randomized Control Trials (RCTs)

In an RCT a treatment X is randomly assigned to some individuals (treatment group), but not to others (control group). If there are differences in an outcome variable Y, they are attributed to the treatment (intervention) and one can claim that "X causes Y". RCTs are an interventional approach and thereby enable deriving cause-effect relationships (**Box 1**). This is supported by the crucial aspect of randomization of group assignment in RCTs. Randomization rules out influential factors on the outcome Y beside the treatment of interest X. Speaking in the terminology of DAGs and the do-operator, randomization erases all arrows that come into X and thereby prevents information about Y from flowing in the non-causal direction[36]. Conveniently, randomization even controls for confounders that cannot be observed, measured or named. This makes randomization not only an effective tool to erase confounding effects but makes RCTs in medical experimentation often seen as the *gold standard* for cause-effect investigations.

168

Box 3 – Types of 3rd variables and associated biases

When investigating the relationship between a predictor (feature) X and an outcome (target) Y, a 3rd variable Z can be related to X and Y in different ways. The different natures can be best visualized by using directed acyclic graphs (DAGs) (**Box 2**).

Confounding

Confounding can be expressed in the form of a causal diagram or formal language (**Box 2**). A confounder Z is a (direct or indirect) common cause of the feature (predictor) X and the target (outcome) Y.

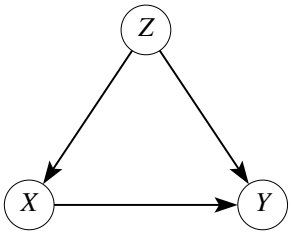

Fig. B3.1 DAG of a confounder Z.

Formally, a confounder can be defined as a variable Z that leads to a discrepancy between the conditional probability of Y given X (*seeing*) and the probability when intervening on X (*doing*):

$$P(Y \mid X) \neq P(Y \mid do(X)) \quad \text{(B3.1)}$$

In the lifespan-drug example from **Box 2** that means that one must ensure that the observed change in Lifespan L is due to the drug itself (do(D)) and is not confounded with other factors Z that tend to shorten or lengthen life. If, instead of intervening, the patient had decided by themselves whether to take the drug (P(L | D)), those other factors Z might influence their decision and lifespan differences between taking and not taking the drug would no longer be solely due to the drug.

Not controlling for a confounder will obscure the causal effect of X on Y. One can either control for the confounder itself or any variable that lies on the path X ← Z → Y.

Other variable types that can lead to biases: Collider – Mediator – Proxy

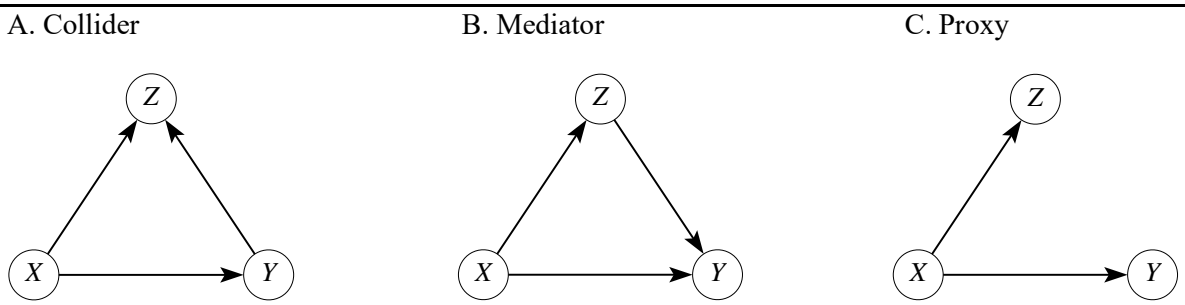

Fig. B3.2 DAG of a collider (A), a mediator (B) and a proxy (C).

A **collider Z** is the common effect of a predictor X and a target Y. Conditioning on a collider induces a spurious (i.e. non-causal) association between X and Y (Berkson's paradox)[41–43] (Fig. B2.2A). In other words, if X and Y were independent to begin with, conditioning on Z will make them dependent (see Berkson's paradox for an example).

A **mediator Z** is caused by X and is a cause of Y[44–46]. For example blood pressure might mediate the relationship between a drug and the risk for a heart attack such that the drug decreases the risk for a heart attack via lowering blood pressure. When interested in the total effect of the predictor on the outcome (X → Y and X → Z → Y), conditioning on Z blocks the causal path X → Z → Y and will hence only reveal a partial effect. When only interested in the direct effect X → Y conditioning on Z can nonetheless lead to biased estimates, if the mediator and the outcome share a common cause because then the mediator is a collider for the predictor and this common cause.

A **proxy Z** is caused by X but has no causal relation to Y[39]. If the predictor X is a perfectly reliable measure of the construct of interest, then controlling for a proxy will not affect the path X → Y. However, in many disciplines X is an unreliable measure of the true causal variable, e.g. a MRI scan for the underlying morphology. In this case, the proxy is a second unreliable measure of the same true predictor (e.g. morphology) and conditioning on this proxy will partition the true predictive effect between the two unreliable proxies so that neither of the unreliable measures will capture the full causal effect[29].

Note: Defining confounding via correlations and not as a causal note is not sufficient because each of the causal structures A.-C. produces a correlation between Z and both X (predictor) and Y (outcome), which could all produce the same correlation matrix. Consequently, correlations cannot help to distinguish between a confounder, a collider or a mediator[47,48].

Types of biases (paradoxes) associated with 3rd variables

Simpson's paradox (confounder bias)

Simpson's paradox is a statistical phenomenon in which the statistical relationship between two variables in a population can appear, disappear, or reverse when splitting the population in subgroups or when aggregating two heterogenous subgroups into a population. For example, two variables might be positively associated in the overall population but either not or negatively associated within the subgroups[49]. More generally, it is characterized by the statistical results of the subgroups differing from the aggregated population. It alerts to cases where at least one of the statistical trends (either in the aggregated data, the partitioned data, or both) cannot represent the causal effects[36].

Berkson's paradox (collider bias)

Berkson's paradox is the opposite of the Simpson's paradox, i.e. it occurs when falsely conditioning on a variable that is the effect of both the feature(s) and the target (collider). Conditioning on such a collider creates a spurious association between the feature(s) and the target. For example, performing a study on patients who are hospitalized, one controls for/conditions on hospitalization. However, if only a disease

1 and a disease 2 together could lead to hospitalization in the first place (with no causal relation between the diseases), conditioning on hospitalization (by performing the study only on hospitalized patients) would introduce non-existing relation between disease 1 and disease 2.

To answer questions of *why* with predictive models, it is not enough to consider associative (correlative) patterns in the data. It is essential to additionally acknowledge directionalities, i.e. the cause-effect structure between relevant variables. Among such relevant variables, it is crucial to carefully distinguish between different types of 3rd variables to identify confounders (**Box 3**). Considering either a standard set of variables or *every conceivable* 3rd variable as a confounder can lead to biased models if these variables in fact were a mediator, collider or proxy. Through randomization, RCTs disable a correct set of confounders without introducing new confounders and thereby implicitly have a causal reasoning integrated by design (**Box 2**). However, RCTs are often not feasible for a variety of reasons, such as ethical concerns, interest in population-based insights or individual-level predictions (precision medicine). To achieve RCT-like causal insights with predictive models (**Box 1** rung 2), one must find a means to purposefully integrate causal reasoning when building predictive models.

3. A 5-step approach to identify valid deconfounders in causal predictive modelling

The core mechanism for integrating causal reasoning into a predictive model is through the identification of and adjustment for a correct set of deconfounders (see **Box 4** for definition). Identifying such a correct set of deconfounders requires a causal analysis around the cause-effect relationship of interest. This causal analysis relies on domain knowledge about the process that generates the observed data. The causal analysis results in a causal diagram or directed acyclic graph (DAG) (**Box 2**), which allows to identify different possibilities for confounder adjustment. For easy transferability to any kind of research project, we in the following describe a 5-step approach to identify a correct set of deconfounders (**Fig. 2**). We exemplify each theoretical step with the previously introduced neuroimaging GMV-HGS prediction example.

3.1. Step 1 and step 2 - Prerequisites and the causal question

In step 1, the *general predictive aim*, such as the out of sample (OOS) prediction of HGS from GMV, is the basis for formulating the *causal aim*, for example if GMV causes changes in HGS (**Fig. 2**, step 1).

In step 2, the causal question refines the causal aim by adding more detailed as well as interventional or counterfactual assumptions (**Fig. 2**, step 2). The causal question expresses the interest in the direct cause-effect relationship of a feature (predictor) X on a target (outcome) Y, i.e. X → Y. For example, if an individual managed to increase their GMV, would that make their hand grip stronger (GMV → HGS)? This causal question requires a causal predictive model because neither a direct interventional approach such as an RCT nor a counterfactual approach is possible (**Box 1**). One cannot experimentally manipulate the volume of grey matter of a participant with the hope to observe if this manipulation of volume will lead to changes in HGS (interventional). Even less, it is not possible that one could change GMV and not change the GMV at the same time in the same individual (counterfactual). Additionally, one might be interested in individual-level, i.e. out of sample predictions. We here focus on an interventional causal predictive model.

3.2. Step 3 - Performing the causal analysis to build a causal diagram and identify deconfounders

Step 3 consists of creating a DAG around the hypothesized direct cause-effect relationship X → Y. The DAG is built through a causal analysis that determines influential factors on both X and Y. The causal analysis starts off by asking about known and conceivable causes of the target Y and then repeats the question for subsequently added 3rd variables (**Fig. 2**, step 3). The answers can be found from previous research and rely on domain-expert knowledge that translates into cause-effect *arrow*-information. This procedure creates a DAG in a bottom-up way. For example, known direct causes of HGS could be *lower arm/upper body muscle mass* (muscle mass → HGS) and the muscles' supply of oxygen and nutrients (oxygen supply → HGS). Additionally, GMV is the conceivable cause of HGS to be investigated (GMV → HGS). In the next iteration, known or conceivable causes of *muscle mass* could be *sex hormones*, *eating behaviour*, *strength training*, *age* etc.. The feature GMV is influenced by *TIV*, age, *sex hormones* and further - potentially unmeasurable or

unobserved - environmental and behavioural factors. This bottom-up procedure is continued until the DAG contains enough information to determine a suitable deconfounding strategy[1] (**Fig. 3**a, **Box 4**).

There are three essential aspects to keep in mind, when building a DAG. First, the feature must be included as a potential cause of the target as this is the goal of the causal question. Second, an omitted arrow restricts the assumed causal effect to zero, while a present arrow remains agnostic about the magnitude of the causal effect of the causal effect. Therefore, not putting an arrow can imply a more precise statement about the cause-effect relationship of two variables than putting an arrow. Third, the DAG relies on established cause-effect relationships but can, and oftentimes must, also be based on ambiguous cause-effect assumptions, for example when there is not yet enough existing causal domain-knowledge. Consequently, there can exist several DAGs for the same causal question. Even though it can be ambiguous, performing a causal analysis is nonetheless beneficial. Formalizing made assumptions with a DAG enables transparent communication. Additionally, the DAG provides a basis for interpreting the resulting "provisional causal"[36] insights gained by the predictive model. Provisional causality thereby means causality contingent upon the set of assumptions that the DAG advertises. Eventually, the causal analysis forces a researcher to precisely *think* about the to-be-answered question and to formalize the causal assumptions.

In contrast to a pure correlative analysis, the DAG reveals the distinction of confounding pathways from colliders and mediators (**Box 3**) and enables a variety of deconfounding strategies (**Box 4**). Correct deconfounding thereby is *the* means to allow predictive models to give provisional causal insights. A correct set of deconfounders can be identified from a DAG either by following the graph rules or by employing available (online) tools (e.g. DAGitty[50] or CausalFusion (https://causalfusion.net)). The DAG for the GMV-HGS example contains 10 confounding, i.e. non-causal pathways (**Fig. 3**b red arrows). According to the "backdoor criterion"[36] (**Box 4**) all 10 non-causal pathways between GMV and HGS can be blocked when adjusting for the deconfounders *sex-hormone levels* and *age*.

In the selection of confounders for predictive modelling it can be challenging to know when and if **all** relevant confounders were identified. This uncertainty can be solved through the concept of deconfounders – in contrast to confounders – in combination with the suggested bottom-up causal analysis. Building the DAG bottom-up allows to identify the point where adding more variables to the DAG does not add more useful information. This point is reached when a set of sufficient deconfounders, blocking the non-causal pathways, can be identified. For example, specifying U_2 more precisely in the DAG illustrated in **Fig. 3**a would not give any information gain with respect to the causal question of interest (GMV → HGS).

[1] *Note*: Strategy here does not refer to the kind of statistical tool to use to correct for confounding signals, e.g. linear regression, but to different ways of choosing a right set of deconfounders based on a DAG as described in **Box 4**.

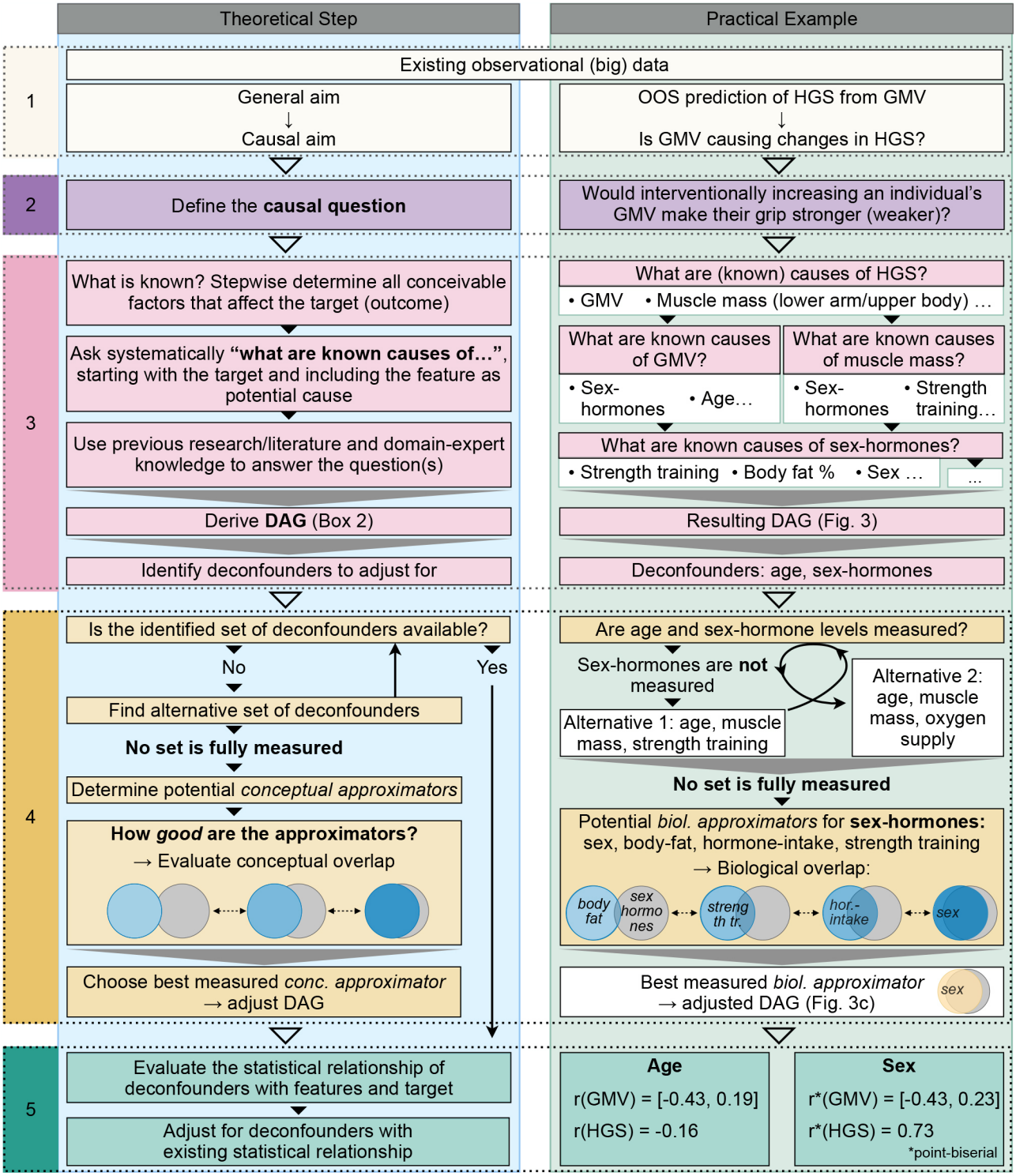

Fig. 2. 5-step recipe for confounder adjustment in causal predictive modelling. The five steps include prerequisites (0), definition of the causal question (1), causal analysis resulting in a DAG and the definition of a minimal set of deconfounders (2), if needed the identification of an alternative set of deconfounders or the replacement of a deconfounder with a conceptual approximator (3) and the statistical evaluation of the identified deconfounders (4). Each step is explained theoretically (left) and by means of the example prediction of hand grip strength (HGS) from grey matter volume (GMV).

3.3. Step 4 - Strategies if not all identified deconfounders are available

When relying on existing observational (big) data, some of the identified deconfounders might not be available, for example *sex-hormone levels* in the example prediction. In such a case, in step 4 the DAG can either reveal an alternative sufficient set of deconfounders for which all variables are measured or the "frontdoor criterion"[36] can be applied (**Box 4**). The DAG for the GMV-HGS example (**Fig. 3**b) reveals that there is neither a fully measured alternative set of deconfounders available (**Fig. 2**, step 4) nor can the frontdoor criterion be applied. The latter can easily happen in the field of neurobiomedicine, where it can be challenging to almost impossible to find a deconfounding variable Z that unambiguously fulfils all three necessary criteria for the frontdoor criterion (**Box 4**). For example, given a neuroimaging derived feature such as GMV, there is no variable for which it can be unambiguously said that it is caused by GMV, and *only* GMV, and that at the same time is a cause of HGS (criterion a and b). The underlying reason is the multi-dimensionality of neurobiomedical phenomena.

The third alternative option in step 4 is the use of a *conceptual approximator*, which makes use of the aforementioned multi-dimensionality. Different neurobiomedical measurements might measure a somewhat similar underlying biological concept, yet still different aspects thereof. For example, *sex* and *sex hormones* express almost the same underlying biology but are nonetheless different measures that also contain non-overlapping biological information: Strength training influences sex-hormone levels[51,52] or body-fat tissue can be hormonally active[53] but neither of them changes the biological sex. One can make use of this multi-dimensionality and overlap in biological information to identify a *biological approximator* as replacement for an unmeasured deconfounder. For instance, here *sex* is the best suited *biological approximator* for the unmeasured deconfounder *sex hormone levels*. The causes of the unmeasured deconfounder thereby serve as candidate *conceptual (here: biological) approximators* which are evaluated based on their conceptual (here: biological) information overlap with the unmeasured variable (**Fig. 2**, step 4).

It is important to note that the *conceptual approximator* cannot directly replace the original deconfounder. The DAG must be modified after replacement because the *conceptual approximator* can change the previously determined cause-effect structure. For example, *strength training* can influence *hormone levels*, but it will not affect the *sex* of a person, which leads to a change in the structure of the DAG (**Fig. 3**c). After modification there are two pathways confounding GMV→HGS (**Fig. 3**c). They can either be blocked by adjusting for *age* and *sex* or by adjusting for *age*, *muscle mass* and *oxygen supply*/*strength training*. This makes *age* and *sex* the minimum correct set of deconfounders for which all variables are measured under the usage of a *biological approximator*.

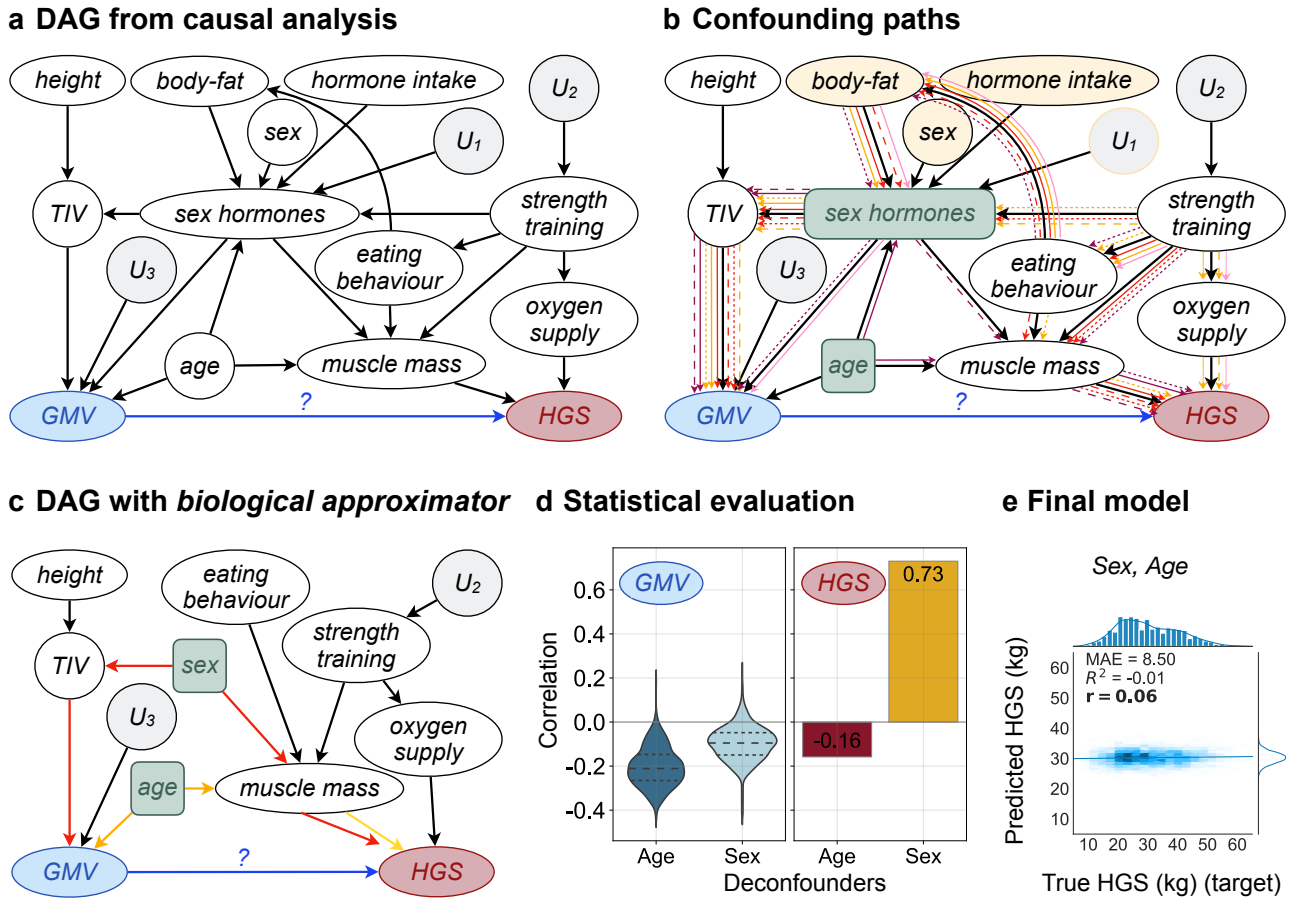

Fig. 3. Practical illustration of 5-step approach with the GMV → HGS predictive example. a) Resulting DAG from a causal analysis. b) *Sex hormones* and *age* (green) qualify as minimum set of correct deconfounders to block all 10 non-causal (confounding) pathways (red arrows) between GMV and HGS. The causes of *sex hormones* (yellow) qualify as candidate *conceptual approximators* for the unmeasured deconfounder.

Box 4 – Ways to account for confounding influences based on DAGs.

Once the underlying causal structure of a causal question was specified in the form of a DAG, there are three ways to identify and account for confounding influences.

Backdoor criterion

A backdoor path is any path from X to Y that starts with an arrow pointing into X, for example X ← Z → Y in Fig. B4.1A. Backdoor paths are *non-causal* paths. To deconfound X and Y one needs to block every *non-causal* path between X and Y without blocking or perturbing causal paths. This can be achieved by adjusting for variables with an incoming arrow into X based on the respective DAG (Fig. B4.1A). These variables are called deconfounders and can differ from the set of confounders. Confounders and deconfounders can differ for example when a confounder does not need to be controlled for because the backdoor path is already blocked by a collider (Fig. B4.1B) or when the actual confounder is unmeasured, but the backdoor path can be blocked by controlling for a measured deconfounder (Fig. B4.1C). RCTs block the non-causal pathways through randomization. Therefore, when picking the right set of deconfounders, statistical adjustment has the same effect as the randomization in RCTs.

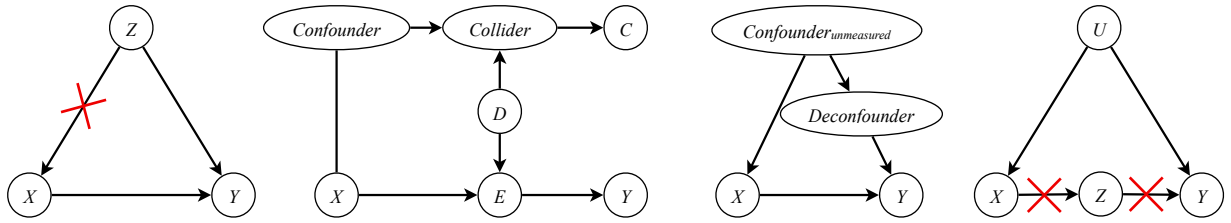

Fig. B4.1 Different positions of a confounder in the DAG structure require different confounder adjustment strategies.

Frontdoor criterion

The backdoor criterion is not feasible when one (or all) deconfounders cannot be measured or are not available. In this case the frontdoor criterion can be applied. The frontdoor criterion requires a variable Z that a) intercepts all direct paths from X → Y, b) there is no backdoor path from X to Z and c) all backdoor paths from Z to Y are blocked by X (Fig. B4.1D). As the relationship between X and Y is confounded by the unobserved variable U, in the frontdoor criterion the effect of X on Y is estimated indirectly by combining the estimate of effect X → Z and of Z → Y.

Instrumental variables

Instrumental variables are similar to a coin flip, which simulates a variable with no incoming arrow, such as Z in Fig. B4.2. Z there qualifies as an instrumental variable because:

a) U and Z are independent, i.e. there is no arrow between U and Z
b) There is an arrow between Z and X
c) There is no direct arrow between Z and Y, i.e. no direct causal connection

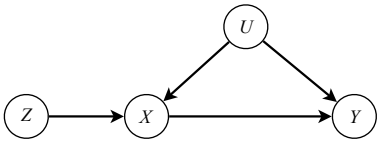

Fig. B4.2 DAG of instrumental variable Z

There is no confounders of the relation between Z and Y, so that any observed association must be causal. Likewise, since the effect of Z on Y goes through X, one can conclude that the observed association between X and Y must also be causal.

3.4. Step 5 - Combining causal and statistical information

Confounders are often defined as any variable that correlates with the feature (predictor) X and the target (outcome) Y. In line with previous works (e.g.[29,36,39,48]), we challenged this *pure* statistical note on confounding and highlighted that correlation does not imply causation. However, causation only becomes relevant when a correlation exists (**Box 1**). We nonetheless put the causal analysis first because this allows for a bottom-up identification of deconfounders in contrast to a top-down pre-definition of confounders that creates insecurities about what confounders to include. By evaluating the statistical relationship of the causally, i.e. knowledge-derived deconfounders with X and Y, step 5 combines the associative (statistical) and the causal perspective (**Fig. 2**, step 5).

Only deconfounders which are statistically and causally relevant must be adjusted for in the predictive model (**Fig. 4**), in our scenario *sex* and *age* (**Fig. 3**d). The threshold for what is considered a sufficient statistical association should be determined based on knowledge about typical association strengths in the domain and is

a question with no strict answer, comparable to other statistical approaches where conventional thresholds are set (e.g. p=0.05 in null hypothesis testing). Deconfounders with no statistical relationship should not be adjusted for as they are in the best case irrelevant but in the worst case would bias the model by leaking information from the deconfounder into the feature or target in the adjustment process[54] (**Fig. 4**). Importantly, it is also indispensable that there is a statistical association between X and Y (**Fig. 1**b). Otherwise it is not meaningful to evaluate if this statistical association (e.g. predictability) does imply a causal connection through the presented causal deconfounding approach. Deconfounders with a statistical but no causal connection are irrelevant because they would not have become part of the set of valid deconfounders based on the causal analysis (**Fig. 4**). For example *length of working week in the main job* correlates with HGS by $r_{Pearson} = -0.24$ and with parcellated GMV in a range of $r_{Pearson} = [-0.11, 0.09]$ but is not part of the DAG. Adjusting the predictive model with the final set of deconfounders leads to a model that allows for provisional causal insights. The *sex* and *age* adjusted GMV→HGS model therefore allows for the provisionally causal insights that GMV is no linear[2] causal predictor of HGS (**Fig. 3**e, see supplementary materials for methods).

Variables with …	Causal relation - no	Causal relation - yes
Statistical relation - yes	… will **not** appear **in DAG** → do not adjust for	Is the variable in the set of deconfounders? **Yes** → Adjust for
Statistical relation - no	… are **irrelevant** for investigating the causal aim → do not adjust for	… have **no shared signal** → do not adjust for

Fig. 4. Combining causal and statistical information on deconfounders can technically lead to four different scenarios. Only deconfounders that play a statistical and causal role for the predictive aim should be adjusted for.

4. Discussion

Predictive neurobiomedical models suffer from a lack of generalizability and replicability fueled by a race for high performance models. We here argue that the field needs more models that aim to deepen the understanding of causal neurobiomedical mechanisms. A key mechanism to achieve causal predictive models is correct confounder adjustment. We proposed a 5-step approach to correctly identify deconfounders based on the combination of causal and statistical investigations. Incorporating causal domain knowledge (possibly supported by large language models, i.e. non-ML AI) about the data generation process into a machine learning model by means of a causal analysis enables to obtain enriched models beyond what is possible by purely data-driven approaches. Proper deconfounding thereby enhances the understanding of biomedical mechanisms and supports the building of reliable (and trustworthy) medical AI tools.

The causal analysis forms the core of causal predictive modelling. If a complete DAG can be derived from this causal analysis, employing methods for deconfounding are straight forward (**Box 4**). However, the multi-dimensionality of neurobiomedical processes can impede the generation of a complete and unambiguous causal diagram. Given that the DAG originates from best knowledge based on literature and domain expertise, this ambiguity makes the decision on a final structure of the DAG to some extent subjective. The causal assumptions nonetheless enable causal insights through targeted confounder control. Those causal insights may be labelled provisional causality because they hold true under the set of assumptions expressed in the DAG. The remaining uncertainty for the causal claims remains as high as the possibility that further existing causal relationships and confounders were not considered. Even though provisional – by adding some causality to the system one gains more causal insights.

A model that provides provisional causality supports answering questions of *why*, which is arguably the ultimate goal of most neurobiomedical investigations. For example, why does person A have a higher HGS

[2] Usage of a linear algorithm for the prediction (see methods in supplementary materials).

than person B? The DAG in **Fig. 3**a represents the known and assumed answers to this question based on literature. For example, person A has a higher HGS because person A has more muscle mass. The results from the provisional causal model inform us that given a linear model (**Fig. 3**e), no variance (R^2 = -0.01) in HGS can be explained by changes in GMV. This means, that based on the assumed DAG and under the employment of a *biological approximator*, GMV is no linear[3] cause of HGS, i.e. person A is not stronger because of a higher/lower volume in grey matter. Such a result may be disappointing: One must perform a timely causal analysis that is potentially incomplete and subjective, the determined confounders are *sex* and *age*, for which one might have adjusted anyway, and the resulting model performance may be lower than hoped for. However, there is an important gain from the investigation: A deeper understanding of a neurobiomedical mechanism.

The causal analysis can be incomplete and to some extent ambiguous, but it allows for the formalization of assumptions and enables the skilful interrogation of nature. Knowing the set of assumptions behind a prediction through such formalization is not less valuable than attempting to circumvent those assumptions with an empirical interventional approach such as an RCT. Additionally, an uncertain answer to the right question is more helpful than a highly certain answer to the wrong question. The causal analysis is motivated by building models that help to better understand neurobiomedical mechanisms. It is needed when one aims to build explainable AI that asks to understand nature or models that are generalizable to new settings. An uncertain answer, i.e. a low performing model, that however helps to answer questions of *why* of neurobiomedicine, is therefore more helpful in the long-term than a high performing model for which it is ambiguous what it means and what questions it answers. Therefore, the low predictive performance of the provisional causal model should not distract from the fact that one learned something about the original causal question, in the example prediction if interventionally increasing an individual's GMV would make their grip stronger. Assuming there was a small predictability, analysing the feature importances that were driving such a model's predictions would be informative, for example revealing what brain areas' GMV are causing stronger HGS. In contrast, the high performing *vanilla* model from **Fig. 1**a does not allow to derive any conclusions about the cause-effect relationship between GMV and HGS. Therefore, it can be better to learn a small effect about a neurobiomedical mechanism of interest than learning a big effect, which's neurobiomedical meaningfulness however remains unclear and that may lead to false conclusions and misinterpretations.

The causal analysis in combination with the use of a *biological approximator* identified *sex* and *age* as deconfounders to adjust for in the GMV-HGS example. The multi-dimensionality of neurobiomedical mechanisms not only allows for the use of *biological approximators*, but it also explains why *sex* and *age* are the two most commonly considered deconfounders. *Sex* and *age* are comparably robust and biologically well-explainable concepts[4], that are often at the beginning of a chain of cause-effect relationships. Thereby, they exhibit a considerably strong biological overlap with many other variables. This makes them a coarse, but – to a varying degree – valid *biological approximator* for other variables further down a cause-effect chain. Here, *sex* served as a *biological approximator* for *sex hormone levels*. When neglecting more fine-grained cause-effect relationships, *sex* could even serve as a *biological approximator* for e.g. *muscle mass*, because males in average have more muscle mass than females. However, the coarser the *biological approximator*, the more it prevents a fine-grained, in depth investigation of the causal question. A coarse *biological approximator* thereby fosters the use of generic average mechanisms, contradictory to the goal of individualized predictions through predictive modelling. Deconfounding predictive models without justification for *sex* and *age* in many cases may not be entirely wrong. However, if not based on a proper causal analysis, such unjustified adjustment may be imprecise and prevents both, replicability and a finer understanding of biological cause-effect relationships.

For many neurobiomedical mechanisms, clear cause-effect relationships are still unknown. Potentially identical processes are approached and described by different disciplines from varying perspectives. This creates multi-dimensional explanations of potentially low-dimensional neurobiomedical mechanisms. The underlying and unambiguous cause-effect chains and networks however remain poorly understood. To determine the functioning, structure, interplay and dimensionality of neurobiomedicine, the field requires the

[3] Usage of a linear algorithm for the prediction (see methods in supplementary materials).
[4] The important investigation on the (un-)ambiguity of the concepts of biological sex and age is beyond the scope of this manuscript and can therefore be found in-depth elsewhere.

integration of information across disciplines, scales and species. The gained insights from provisional causal models can improve and inform the next formalized causal analyses and causal predictive models, which in turn can provide new provisional causal insights. Such a recursive feedback loop within causal predictive modelling but also across scales and species (e.g. from direct interventional animal research) can iteratively clarify and improve what is known about neurobiomedical cause-effect relationships. Additionally, mutual recursive feedback of causal knowledge from experimental setups can inform the DAGs for observational analyses. The observational causal predictive models in turn can both inform experimental setups about the generalizability of effects and create new provisional causal insights. A recursive feedback mechanism thereby is important to avoid accumulation of errors. Thereby the field as a whole, beyond individual research projects, can contribute to shaping a neurobiomedical causal diagram based on the understanding of underlying mechanisms. Ultimately, this could help disentangle the causal structure of neurobiomedicine and determine orthogonal (independent) biological dimensions.

While the investigation and use of cause-effect relationships is more commonly used in fields such as economics or social sciences, it is so far seldomly applied for correctly deconfounding neurobiomedical predictive models. With the proposed 5-step approach, we hope to provide an easy-to-use standard approach for causal predictive modelling. Through causally motivated deconfounding, we aim to foster the relevance of low performing or even null results models if they originate from a "skilful interrogation of nature". Ultimately, mutual recursive feedback loops of causal insights across disciplines, scales and species can enable the field to disentangle the cause-effect structure of neurobiomedical mechanisms. Through understanding and knowledge, this can facilitate reliable and trustworthy AI as a medical tool.

5. References

1. Berisha V, Krantsevich C, Hahn PR, et al. Digital medicine and the curse of dimensionality. *Npj Digit Med*. 2021;4(1):153. doi:10.1038/s41746-021-00521-5

2. Darcy AM, Louie AK, Roberts LW. Machine Learning and the Profession of Medicine. *JAMA*. 2016;315(6):551. doi:10.1001/jama.2015.18421

3. Jiang F, Jiang Y, Zhi H, et al. Artificial intelligence in healthcare: past, present and future. *Stroke Vasc Neurol*. 2017;2(4):230-243. doi:10.1136/svn-2017-000101

4. Arbabshirani MR, Plis S, Sui J, Calhoun VD. Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. *NeuroImage*. 2017;145:137-165. doi:10.1016/j.neuroimage.2016.02.079

5. Benkarim O, Paquola C, Park B yong, et al. The Cost of Untracked Diversity in Brain-Imaging Prediction. *bioRxiv*. Published online June 2021:34. doi:https://doi.org/10.1101/2021.06.16.448764

6. Pulini AA, Kerr WT, Loo SK, Lenartowicz A. Classification Accuracy of Neuroimaging Biomarkers in Attention-Deficit/Hyperactivity Disorder: Effects of Sample Size and Circular Analysis. *Biol Psychiatry Cogn Neurosci Neuroimaging*. 2019;4(2):108-120. doi:10.1016/j.bpsc.2018.06.003

7. Woo CW, Chang LJ, Lindquist MA, Wager TD. Building better biomarkers: brain models in translational neuroimaging. *Nat Neurosci*. 2017;20(3):365-377. doi:10.1038/nn.4478

8. Heinrichs B, Eickhoff SB. Your evidence? Machine learning algorithms for medical diagnosis and prediction. *Hum Brain Mapp*. 2020;41(6):1435-1444. doi:10.1002/hbm.24886

9. Kapoor S, Narayanan A. Leakage and the Reproducibility Crisis in ML-based Science. Published online July 14, 2022. Accessed January 31, 2023. http://arxiv.org/abs/2207.07048

10. Alfaro-Almagro F, McCarthy P, Afyouni S, et al. Confound modelling in UK Biobank brain imaging☆. Published online 2021:17.

11. Dinga R, Schmaal L, Penninx BWJH, Veltman DJ, Marquand AF. *Controlling for Effects of Confounding Variables on Machine Learning Predictions*. Bioinformatics; 2020. doi:10.1101/2020.08.17.255034

12. Weinberger DR, Radulescu E. Finding the Elusive Psychiatric "Lesion" With 21st-Century Neuroanatomy: A Note of Caution. *Am J Psychiatry*. 2016;173(1):27-33. doi:10.1176/appi.ajp.2015.15060753

13. Rao A, Monteiro JM, Mourao-Miranda J. Predictive modelling using neuroimaging data in the presence of confounds. *NeuroImage*. 2017;150:23-49. doi:10.1016/j.neuroimage.2017.01.066

14. Geerligs L, Tsvetanov KA, Cam-CAN, Henson RN. Challenges in measuring individual differences in functional connectivity using fMRI: The case of healthy aging: Measuring Individual Differences Using fMRI. *Hum Brain Mapp*. 2017;38(8):4125-4156. doi:10.1002/hbm.23653

15. Power JD, Barnes KA, Snyder AZ, Schlaggar BL, Petersen SE. Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*. 2012;59(3):2142-2154. doi:10.1016/j.neuroimage.2011.10.018

16. Satterthwaite TD, Wolf DH, Loughead J, et al. Impact of in-scanner head motion on multiple measures of functional connectivity: Relevance for studies of neurodevelopment in youth. *NeuroImage*. 2012;60(1):623-632. doi:10.1016/j.neuroimage.2011.12.063

17. Spisak T. Statistical quantification of confounding bias in predictive modelling. Published online November 1, 2021. Accessed January 31, 2023. http://arxiv.org/abs/2111.00814

18. Bugg JM, Zook NA, DeLosh EL, Davalos DB, Davis HP. Age differences in fluid intelligence: Contributions of general slowing and frontal decline. *Brain Cogn*. 2006;62(1):9-16. doi:10.1016/j.bandc.2006.02.006

19. Hartshorne JK, Germine LT. When Does Cognitive Functioning Peak? The Asynchronous Rise and Fall of Different Cognitive Abilities Across the Life Span. *Psychol Sci*. 2015;26(4):433-443. doi:10.1177/0956797614567339

20. Horn (1967) - age differences in fluid and crystallized intelligence.pdf.

21. Kahlert J, Gribsholt SB, Gammelager H, Dekkers OM, Luta G. Control of confounding in the analysis phase – an overview for clinicians. *Clin Epidemiol*. 2017;Volume 9:195-204. doi:10.2147/CLEP.S129886

22. Rao A, Monteiro JM, Ashburner J, et al. A comparison of strategies for incorporating nuisance variables into predictive neuroimaging models. In: *2015 International Workshop on Pattern Recognition in Neuroimaging*. ; 2015:61-64.

23. Kostro D, Abdulkadir A, Durr A, et al. Correction of inter-scanner and within-subject variance in structural MRI based automated diagnosing. *NeuroImage*. 2014;98:405-415. doi:10.1016/j.neuroimage.2014.04.057

24. Abdulkadir A, Ronneberger O, Tabrizi SJ, Klöppel S. Reduction of confounding effects with voxel-wise Gaussian process regression in structural MRI. In: *2014 International Workshop on Pattern Recognition in Neuroimaging*. IEEE; 2014:1-4.

25. Dukart J, Schroeter ML, Mueller K, The Alzheimer's Disease Neuroimaging Initiative. Age Correction in Dementia – Matching to a Healthy Brain. Valdes-Sosa PA, ed. *PLoS ONE*. 2011;6(7):e22193. doi:10.1371/journal.pone.0022193

26. Snoek L, Miletić S, Scholte HS. How to control for confounds in decoding analyses of neuroimaging data. *NeuroImage*. 2019;184:741-760. doi:10.1016/j.neuroimage.2018.09.074

27. Miller KL, Alfaro-Almagro F, Bangerter NK, et al. Multimodal population brain imaging in the UK Biobank prospective epidemiological study. *Nat Neurosci*. 2016;19(11):1523-1536. doi:10.1038/nn.4393

28. Weinstein SM, Davatzikos C, Doshi J, Linn KA, Shinohara RT, For the Alzheimer's Disease Neuroimaging Initiative. Penalized decomposition using residuals (PeDecURe) for feature extraction in the presence of nuisance variables. *Biostatistics*. Published online August 11, 2022:kxac031. doi:10.1093/biostatistics/kxac031

29. Wysocki AC, Lawson KM, Rhemtulla M. Statistical Control Requires Causal Justification. *Advances in Methods and Practices in Psychological Science*. 2022;5(2).

30. Becker TE. Potential Problems in the Statistical Control of Variables in Organizational Research: A Qualitative Analysis With Recommendations. *Organ Res Methods*. 2005;8(3):274-289. doi:10.1177/1094428105278021

31. Bernerth JB, Aguinis H. A Critical Review and Best-Practice Recommendations for Control Variable Usage. *Pers Psychol*. 2016;69(1):229-283. doi:10.1111/peps.12103

32. Atinc G, Simmering MJ, Kroll MJ. Control Variable Use and Reporting in Macro and Micro Management Research. *Organ Res Methods*. 2012;15(1):57-74. doi:10.1177/1094428110397773

33. Carlson KD, Wu J. The Illusion of Statistical Control: Control Variable Practice in Management Research. *Organ Res Methods*. 2012;15(3):413-435. doi:10.1177/1094428111428817

34. Pourhoseingholi MA, Baghestani AR, Vahedi M. How to control confounding effects by statistical analysis. *Gastroenterol Hepatol Bed Bench*. 2012;5(2):79-83.

35. Chyzhyk D, Varoquaux G, Milham M, Thirion B. How to remove or control confounds in predictive models, with applications to brain biomarkers. *GigaScience*. 2022;11:giac014. doi:10.1093/gigascience/giac014

36. Pearl J, Mackenzie D. *The Book of Why: The New Science of Cause and Effect*. Basic Books; 2018.

37. Quinonero-Candela J, Sugiyama M, Schwaighofer A, Lawrence ND. *Dataset Shift in Machine Learning*. Mit Press; 2008.

38. Huyen C. *Designing Machine Learning Systems: An Iterative Process for Production-Ready Applications*. First edition. O'Reilly Media, Inc; 2022.

39. Pearl J. Causal inference in statistics: An overview. *Stat Surv*. 2009;3(none). doi:10.1214/09-SS057

40. Sudlow C, Gallacher J, Allen N, et al. UK Biobank: An Open Access Resource for Identifying the Causes of a Wide Range of Complex Diseases of Middle and Old Age. *PLOS Med*. 2015;12(3):e1001779. doi:10.1371/journal.pmed.1001779

41. Elwert F, Winship C. Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annu Rev Sociol*. 2014;40(1):31-53. doi:10.1146/annurev-soc-071913-043455

42. Rohrer JM. Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data.

43. Berkson J. Limitations of the Application of Fourfold Table Analysis to Hospital Data. *Biom Bull*. 1946;2(3):47. doi:10.2307/3002000

44. Baron RM, Kenny DA. The moderator–mediator variable distinction in social psychological research: Conceptual, strategic, and statistical considerations. *J Pers Soc Psychol*. 1986;51(6):1173.

45. Hayes AF. Beyond Baron and Kenny: Statistical mediation analysis in the new millennium. *Commun Monogr*. 2009;76(4):408-420.

46. Judd CM, Kenny DA. Process analysis: Estimating mediation in treatment evaluations. *Eval Rev*. 1981;5(5):602-619.

47. Maxwell SE, Cole DA. Bias in cross-sectional analyses of longitudinal mediation. *Psychol Methods*. 2007;12(1):23-44. doi:10.1037/1082-989X.12.1.23

48. Pearl J. Causal diagrams for empirical research.

49. Sprenger J, Weinberger N. Simpson's paradox. In: Zalta EN, ed. *The Stanford Encyclopedia of Philosophy*. Summer 2021. Metaphysics Research Lab, Stanford University; 2021. https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/

519 50. Textor J, Hardt J, Knüppel S. DAGitty: A Graphical Tool for Analyzing Causal Diagrams. *Epidemiology*.
520 2011;22(5):745. doi:10.1097/EDE.0b013e318225c2be

521 51. Swain CTV, Drummond AE, Boing L, et al. Linking Physical Activity to Breast Cancer via Sex Hormones,
522 Part 1: The Effect of Physical Activity on Sex Steroid Hormones. *Cancer Epidemiol Biomarkers Prev*.
523 2022;31(1):16-27. doi:10.1158/1055-9965.EPI-21-0437

524 52. Ambroży T, Rydzik Ł, Obmiński Z, et al. The Effect of High-Intensity Interval Training Periods on
525 Morning Serum Testosterone and Cortisol Levels and Physical Fitness in Men Aged 35–40 Years. *J Clin*
526 *Med*. 2021;10(10):2143. doi:10.3390/jcm10102143

527 53. Tchernof A, Després JP, Bélanger A, et al. Reduced testosterone and adrenal C19 steroid levels in obese
528 men. *Metabolism*. 1995;44(4):513-519. doi:10.1016/0026-0495(95)90060-8

529 54. Hamdan S, Love BC, von Polier GG, et al. Confound-leakage: confound removal in machine learning
530 leads to leakage. *GigaScience*. 2023;12.

531

Acknowledgments

This research has been conducted using the UK Biobank Resource under application number 41655. This research was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 431549029 - Collaborative Research Centre CRC1451 on motor performance project B05.

We additionally want to acknowledge Prof Dr Bert Heinrich and Dr Jan-Hendrik Heinrichs from the group for Neuroethics and Ethics in AI at the INM-7, Research Centre Juelich, Germany, for their truly valuable and inspiring contribution to the manuscript from a philosophical perspective on AI.

Author contributions

All authors contributed to discussing, reviewing, and editing the content of the manuscript, and agreed to the final version of the manuscript.

Competing interests

The authors declare no competing interests.