TECHNICAL REPORT
R-20
October, 2025

SENSITIVITY ANALYSIS OF CAUSAL EFFECTS UNDER IMPERFECTLY OBSERVED BINARY CONFOUNDING: APPLICATION TO THE OBESITY PARADOX

By Drago Plecko[1,a],

[1]Department of Statistics & Data Science, UCLA, [a]drago@stat.ucla.edu

Measurement error is a common issue in many statistical applications, and can be particularly difficult to handle when the error pattern is informative. In this manuscript, we consider estimation of average treatment effects (ATE) under imperfectly observed confounding. In particular, we assume that binary confounders may be reported incorrectly, with the data containing a biased proxy of the true confounders. In a sensitivity-type analysis, we answer the question: which rate of (informative) measurement error would be needed to change the sign of the causal effect estimate? The interpretable sensitivity parameter is the data *fidelity* – the probability that a confounder is recorded correctly. We apply the analysis to the obesity paradox, a well-known epidemiological phenomenon in which obese critically ill patients show better survival rates in the intensive care unit than their leaner counterparts. This is a paradox since obesity is usually associated with health risks and increased long-term all-cause mortality. Common explanations of this paradox are through confounding by comorbidities such as cancer or smoking, that reduce patients' weight and increase the mortality risk. However, such comorbidities are recorded imperfectly in electronic health records, and biased reporting of binary comorbidities needs to be considered explicitly.

1. Introduction.

1.1. *Measurement Error.* Measurement error is common in almost all applied sciences, and this is witnessed by the volume of empirical methods for handling measurement error (Fuller, 2009; Carroll et al., 2006; Buonaccorsi, 2010; Rothman et al., 2008). Depending on the pattern of measurement error, the analysis may be more or less difficult to perform. In this context, the literature distinguishes between non-differential and differential errors (Carroll et al., 2006). Non-differential errors are those for which the imperfectly measured values of the covariates do not depend on the outcome, given the true values of the covariates. Otherwise, differential errors are to said to occur, meaning that the outcome values may be informative even if the true covariate values were observed. Consequences of measurement error range from loss in power or reduced effective sample size (often the case with non-differential errors) to complete infeasibility of answering the study question (which may happen with differential errors).

1.2. *Causal Inference.* Causal inference is the field of study concerned with inference of cause and effect relations among variables. Causal effect inference from observational data is commonly performed under suitable assumptions. In the graphical approach to causality (Pearl, 2000), such assumptions are encoded through an object known as the causal diagram. Inference of effects is then licensed by different graphical criteria, such as the back-door criterion (Pearl, 2000). Interestingly, assumptions on measurement error can also be encoded

Keywords and phrases: Causal Inference, Graphical Models, Proximal Inference, Measurement Error.

in graphical models (Kuroki and Pearl, 2014), and we deploy a graphical formalism in this manuscript. In particular, we consider the estimation of an average treatment effect (ATE, also known as total effect), written $\mathbb{E}[Y \mid do(X = x_1)] - \mathbb{E}[Y \mid do(X = x_0)]$, for an outcome Y and a binary treatment $X \in \{x_0, x_1\}$, in a differential measurement error setting where confounding variables are imperfectly observed. Here, $do(\cdot)$ is the Pearl's do-operator (Pearl, 1995). In the sequel, we describe the application that motivates the methods developed in this manuscript.

1.3. *The Obesity Paradox.* Obesity and overweight are increasingly common in developed countries (World Health Organization, 2023), with an estimated 42% of adults obese in the United States (Centers for Disease Control and Prevention, 2020). Importantly, obesity is linked to numerous comorbidities and risk factors for life-threatening complications (Castro et al., 2014; Haslam et al., 2006; Schelbert, 2009). Furthermore, obesity is also strongly associated with increased long-term all-cause mortality (Berrington de Gonzalez et al., 2010; Flegal et al., 2005), and in chronic medicine there is a widespread consensus on some of the negative effects of obesity. The most commonly used measure of obesity is the *body-mass index* (BMI), defined as the ratio of weight (in kilograms) w and squared height (in meters) h^2, BMI $= w/h^2$. This measure is also considered in this work.

In the context of critical illness, that is for patients admitted to intensive care units (ICUs) after surgical procedures or due to various acute medical complications, the effects of obesity are remarkably different. A large body of evidence demonstrates that for both medical and surgical ICU populations, increased BMI is associated with improved survival (Hutagalung et al., 2011; Mullen et al., 2009; Tremblay and Bandi, 2003; Hainer and Aldhoon-Hainerová, 2013), and even with improved functional outcomes after discharge (Yeo et al., 2023). This widely studied and surprisingly robust epidemiological phenomenon, entirely contrary to the understanding of obesity in chronic medicine, has thus been termed *the obesity paradox*.

Various explanations of the paradox have been proposed. Some causal explanations state that the increased BMI may represent greater physiological reserve in terms of stored protein or energy (Plečko et al., 2021) that is helpful in critical illness. Other causal explanations discuss the protective effect of fat tissue (adipocytes) in reducing inflammatory response (Tilg and Moschen, 2006; Fantuzzi, 2005). However, the most commonly considered explanation is through confounding – decreased BMI represents a state of reduced well-being, due to various comorbidities (such as cancer, smoking, etc.) that confound the BMI relationship with outcome. Interestingly, such arguments are in principle testable from appropriate data by adjusting for this type of confounding. As we demonstrate later on, though, adjusting for known comorbidities does not remove the nexus between obesity and improved outcomes in critical illness.

In this context, considerations of measurement error are of prime importance. Data recorded in electronic health records (EHR) may be subject to various types of measurement error, and this may affect the estimation of the causal effect of BMI on outcome[1]. To this end, we introduce the following type of analysis. We assume that existing comorbidities are recorded imperfectly, with an existing comorbidity not recorded with a probability ϕ, and a non-existing comorbidity recorded falsely with probability ϕ'. Then, we aim to answer the question of what is the smallest value of ϕ (or ϕ') such that the causal effect estimate changes sign compared to the effect estimate for the no measurement error scenario $\phi = \phi' = 0$. We also consider

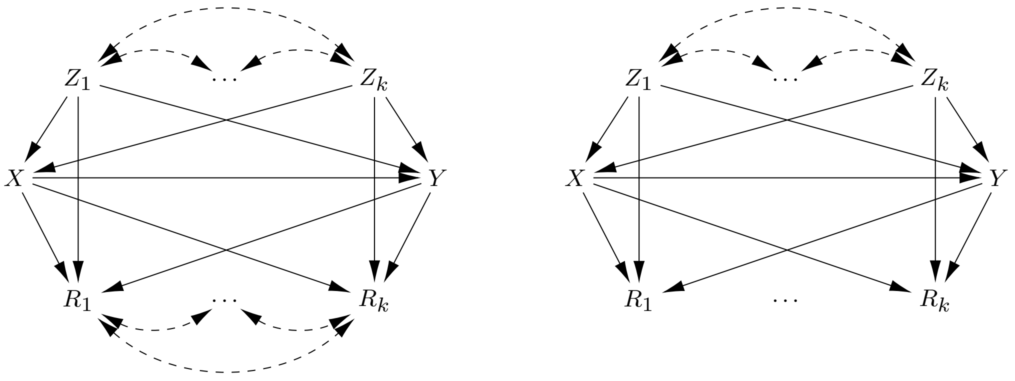
[1] Issues of manipulability of variables are not discussed in this manuscript, for a discussion see (Pearl, 2018). We also refer the reader to a large literature investigating causal effects of BMI on outcome in critical illness (Hainer and Aldhoon-Hainerová, 2013; Hutagalung et al., 2011; Decruyenaere et al., 2020; Banack and Kaufman, 2014).

extensions where the comorbidity measurement error depends on the BMI (i.e., the treatment) and/or death (outcome). We apply our method to a large hospital population from the Beth Isreal Deaconess Medical Center in Boston, Massachusets (MIMIC-IV dataset (Johnson et al., 2020)), and demonstrate that levels of informative measurement error of $< 10\%$ would be sufficient to explain away the obesity paradox (see Sec. 3 for details).

Our work is related to the graphical approach for missing data (Mohan and Pearl, 2021; Nabi et al., 2020), but our setting differs from typical missingness scenarios. In our case, missingness is not directly observed – there is no distinction between a missing value and a recorded negative value of 0, similar to the context discussed in (Dai et al., 2024). As a result, the observed values of confounding variables act as proxies for their true values. In this context, our work is related to and extends some of the ideas in proximal inference (Kuroki and Pearl, 2014; Cui et al., 2024), with initial ideas appearing in the epidemiology literature (Rothman et al., 2008, Ch. 19). Further, we also mention a related graphical approach for sensitivity with discrete data (Duarte et al., 2024). Finally, our work is related to the notion of E-value (VanderWeele and Ding, 2017) that attempts to find the smallest value of a sensitivity parameter that changes the sign of the effect estimate, which is used for sensitivity to unobserved confounding, while our focus is on sensitivity to measurement error. This paper's contributions are:

(i) We frame the problem of inferring causal effects from imperfectly observed binary data (Def. 1), a setting that commonly appears in practice. We prove that the problem can be solved for a wide range of measurement error settings (Thm. 1).
(ii) We propose non-parametric and parametric estimators for the problem of recovering causal effects from imperfect proxies, and provide theoretical guarantees (Thm. 2). We further extend our setting to include correctly observed continuous confounders.
(iii) We introduce the notion of ϕ-value – the minimal measurement error that can change the sign of the estimated causal effect (Def. 3), and give a simple procedure for how it can be computed.
(iv) We apply the developed methodology to analyze the obesity paradox on a large dataset from a tertiary hospital center in the United States, thereby adding to the literature on critical care medicine (see Sec. 3).

1.4. *Preliminaries.* We use the language of structural causal models (SCMs) (Pearl, 2000). An SCM is a tuple $\mathcal{M} := \langle V, U, \mathcal{F}, P(U) \rangle$, where V, U are sets of endogenous (observable) and exogenous (latent) variables, respectively, $\mathcal{F}$ is a set of functions f_{V_m}, one for each $V_m \in V$, where $V_m \leftarrow f_{V_m}(\mathrm{pa}(V_m), U_{V_m})$ for some $\mathrm{pa}(V_m) \subseteq V$ and $U_{V_m} \subseteq U$. The set $\mathrm{pa}(V_m)$ is called the parent set of V_m. $P(U)$ is a strictly positive probability measure over U. Each SCM $\mathcal{M}$ is associated with a causal diagram $\mathcal{G}$ (Bareinboim et al., 2022) over the node set V where $V_m \to V_\ell$ if V_m is an argument of f_{V_ℓ}, and $V_m \dashleftarrow\dashrightarrow V_\ell$ if the corresponding U_{V_m}, U_{V_ℓ} are not independent. Each SCM is also associated with observational and interventional distributions. The observational distribution, labeled $P(V)$, is obtained by sampling the noise variables U from the distribution $P(U)$, and evaluating these variables using the mechanisms in $\mathcal{F}$. An interventional distribution is obtained from a submodel $\mathcal{M}_x$, which is an SCM in which all equations in $\mathcal{F}$ associated with X are replaced by $X = x$. With $P(Y \mid do(X = x))$ we denote the interventional distribution of variables Y in the submodel M_x. We say that an observational distribution $P(Y \mid do(X = x))$ is *identifiable* if it can be uniquely computed from the causal diagram $\mathcal{G}$ and the observational distribution $P(V)$. An extended discussion of the above notions, with motivating examples, is provided in Supplement E.



(a) General diagram.

(b) Independent measurement error diagram.

Fig 1: Causal diagrams of the measurement error setting.

2. Methods.

We now begin by casting the previously described problem of the obesity paradox into the framework of SCMs.

Example (Obesity Paradox). *Patients are admitted to a hospital. At the time of admission, or during the hospital stay, their BMI is recorded. Patients who are underweight (BMI $< 18.5kg/m^2$) or morbidly obese (BMI $\geq 35kg/m^2$) are excluded from our analysis, since these groups are known to have poor outcomes. In our setting, BMI is binarized, $X = \mathbb{1}(BMI \geq 25kg/m^2)$, so that $X = x_1$ indicates overweight or obesity, and $X = x_0$ indicates a BMI in the normal range. Furthermore, a number of chronic comorbidities (such as diabetes, cancer, etc.) that may influence the outcome are also recorded, denoted by $R_1, \dots, R_k$, where every $R_m \in \{0, 1\}$ is binary. However, the process of recording comorbidities may be imperfect, so let the true values of the comorbidities be denoted by $Z_1, \dots, Z_k$. At the end of the hospital stay, patient mortality is recorded (label Y, value 0 for survival, value 1 for in-hospital death). We are interested in inferring the causal effect of X on Y from the observed data on (R, X, Y), written $\mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]$.* □

2.1. *A Model of Measurement Error.*

The assumed causal diagram of the data-generating process described in the example is given in Fig. 1a. The true comorbidities $Z_1, \dots, Z_k$ are connected by bidirected edges, corresponding to the fact that the comorbidities may not be independent, and may be confounded by the patient's latent physiological well-being. The Z_m variables then affect the BMI X (e.g., comorbidities may decrease or increase BMI), and also affect the outcome Y (most comorbidities increase the probability of a negative outcome). Finally, the value of R_m depends on Z_m, but the measurement error pattern may depend on X, Y, encoded in the arrows $X \to R_m, Y \to R_m$. In particular, we assume the following:

Definition 1 (Imperfect Observations). *Let Z_m be a binary confounder, and R_m its recorded value. The mechanism of R_m can be represented as*

$$R_m \leftarrow \begin{cases} \textit{Bernoulli}(\phi'_{xym}) & \textit{if } Z_m = 0 \\ \textit{Bernoulli}(1 - \phi_{xym}) & \textit{if } Z_m = 1, \end{cases} \tag{1}$$

where ϕ'_{xym} is the probability of a false positive ($R_m = 1$, but true $Z_m = 0$), and ϕ_{xym} the probability of a false negative ($R_m = 0$, but true $Z_m = 1$). Values of ϕ'_{xym}, ϕ_{xym} may depend *on the index m and values $X = x, Y = y$. Finally, if*

$$R_m \leq Z_m \ \forall m \text{ almost surely,} \tag{2}$$

we say that the no false positives (NFP) assumption holds, meaning that whenever the confounder is not present ($Z_m = 0$), it will be recorded as not present ($R_m = 0$).

Bidirected arrows between R_m nodes in Fig. 1a imply that measurement errors of confounders may be dependent across different variables $R_1, \ldots, R_k$.

2.2. *Effect Recovery.* As mentioned before, we are interested in recovering the average treatment effect (ATE) $\mathbb{E}[Y \mid do(X = 1)] - \mathbb{E}[Y \mid do(X = 0)]$. Note that the ATE can be written as

$$\mathbb{E}[Y \mid \mathrm{do}(x_1)] - \mathbb{E}[Y \mid \mathrm{do}(x_0)] = \sum_z [P(y \mid x_1, z) - P(y \mid x_0, z)]P(z). \tag{3}$$

Terms $P(z), P(y \mid x, z)$ would be easily obtained from the joint $P(z, x, y)$. Instead, we have access to the joint of (R, X, Y), written $P(r, x, y)$. However, since we can factorize $P(z, x, y) = P(z \mid x, y)P(x, y)$, and $P(x, y)$ can be obtained from $P(r, x, y)$ easily, we may recover the joint $P(z, x, y)$ if we can recover $P(z \mid x, y)$. This task is discussed next.

We first introduce some notation. Let k be the dimensionality of the Z-space. Since each Z_m is binary, the vector Z can take 2^k different values. We index with $z^{(i)}$ the value of the Z that corresponds to the bit-representation of number i, for example $z^{(0)} = (0, \ldots, 0)$, $z^{(1)} = (1, \ldots, 0)$, etc. Furthermore, we denote by $p_{z|xy}$ the vector containing the values of the conditional distribution $P(z \mid x, y)$, i.e., $p_{z|xy} = (P(z^{(0)} \mid x, y), \ldots, P(z^{(2^k-1)} \mid x, y))$, and analogously for $p_{r|xy}$. Further, p_z is defined as the vector of $P(z)$ probabilities, defined through $\sum_{x,y} P(x, y)p_{z|xy}$, while $p_{y|xz}$ is the vector of $P(y \mid x, z)$ probabilities across different $Z = z$ values. Notice that the following identity holds:

$$\underbrace{\begin{pmatrix} \vdots \\ P(r^{(i)} \mid x, y) \\ \vdots \end{pmatrix}}_{p_{r|xy}} = \underbrace{\begin{pmatrix} \ddots & \vdots & \vdots \\ \vdots & P(r^{(i)} \mid z^{(j)}, x, y) & \vdots \\ \vdots & \vdots & \ddots \end{pmatrix}}_{A_{xy}\,(\text{size } 2^k \times 2^k)} \underbrace{\begin{pmatrix} \vdots \\ P(z^{(j)} \mid x, y) \\ \vdots \end{pmatrix}}_{p_{z|xy}} \tag{4}$$

where A_{xy} is the matrix of conditional probabilities $P(r^{(i)} \mid z^{(j)}, x, y)$ that tells us how likely $R = r^{(i)}$ is to be observed given the confounders $Z = z^{(j)}$ and $X = x, Y = y$. We refer to the set of matrices $\{A_{xy}\}_{x,y \in \{0,1\}}$ as the fidelity pattern. In particular, we distinguish four different types of fidelity patterns, depending on how measurement errors are affected by values of X, Y.

Definition 2 (Fidelity Pattern Modification). *Let the matrices $\{A_{xy}\}_{x,y}$ be the fidelity pattern. We say that the fidelity pattern is*

(i) *agnostic if $A_{x_0y_0} = A_{x_1y_0} = A_{x_0y_1} = A_{x_1y_1}$,*
(ii) *treatment-modified or x-modified if $A_{x_0y_0} = A_{x_0y_1}$ and $A_{x_1y_0} = A_{x_1y_1}$,*
(iii) *outcome-modified or y-modified if $A_{x_0y_0} = A_{x_1y_0}$ and $A_{x_0y_1} = A_{x_1y_1}$,*
(iv) *treatment-outcome-modified or xy-modified if none of the above are true.*

Distinguishing between the above pattern types may be helpful in practice, since some types may be more plausible than others, depending on the application.

Going back to the problem of recovering the causal effect of interest, the key intuition is the following. If the matrix A_{xy} is invertible, the conditional $p_{r|xy}$ uniquely determines $p_{z|xy}$. In the following theorem, we establish two relatively general conditions under which the causal effect of interest can be recovered:

THEOREM 1 (Recovery Conditions). *The average treatment effect of X on Y can be uniquely computed from the joint distribution $P(r,x,y)$ provided matrices $\{A_{xy}\}_{x,y\in\{0,1\}}$, if one of the following two conditions holds:*

(a) *The no false positive assumptions holds, $R_m \leq Z_m$ for every component $m \in \{1,\ldots,k\}$, and $P(R = z^{(i)} \mid Z = z^{(i)}, x, y) > 0 \,\forall\, z^{(i)}, x, y$,*
(b) *$P(R = z^{(i)} \mid Z = z^{(i)}, x, y) > \frac{1}{2} \,\forall\, z^{(i)}, x, y$.*

The theorem's proof is provided in Supplement A. The first case covers effect recovery under the no false positive (NFP) assumption, which states that a confounder that is absent will always be recorded as absent. In our context of analyzing electronic health records (EHR) data, such an NFP assumption (Eq. 2) may be seen as plausible. In this case, the effect can be recovered whenever there is a non-zero probability of recording all the comorbidities correctly (for any setting of the values). The second case covers arbitrary fidelity patterns, where even false positives may occur. In this case, the effect can be recovered if, for any setting of the comorbidities Z, we record the correct value of Z with a probability greater than a half. The above theorem shows that the causal effect can be recovered in surprising generality, even when the fidelity pattern depends on both the treatment X and the outcome Y. This result is related to an earlier result of Kuroki and Pearl (2014), but extends the setting substantially, by allowing the fidelity pattern to depend both on the treatment and outcome, and establishing relatively general recovery conditions. Here, we also remark that the approach proposed in this manuscript recovers the true distribution of the confounders $P(z)$, and hence all of the results can be easily translated to other scales apart from ATE, such as risk or odds ratios and conditional average treatment effects.

2.2.1. *Sensitivity Parameter Dimensionality.* As discussed in the sequel, the matrices A_{xy} will play the role of sensitivity parameters, which can be specified by the data analyst. In practice, specifying a full matrix A_{xy} would require specifying 2^{2k} parameters, which may be too expressive for practical purposes when $k > 2$. For this reason, we consider two types of fidelity patterns that simplify the specification of the matrices A_{xy}, through a specification of a smaller number of parameters Φ_{xy} (in such instances, we may sometimes write $A_{xy}(\Phi_{xy})$ to indicate such a parametrization). We emphasize that these fidelity patterns are considered to alleviate the difficulty of specifying the sensitivity parameters, while they are not necessary for any of the inference steps later in the text (Sec. 2.4).

2.2.1.1. Independent Fidelity. The first pattern we consider is that of independent fidelity (IF, for short). Formally, we say independent fidelity holds if

$$P(r \mid z, x, y) = \Pi_{m=1}^{k} P(r_m \mid z_m, x, y), \tag{5}$$

meaning that measurement errors for the component Z_m depend only on X, Y, and the actual Z_m value. In this setting, specifying the parameters

$$\phi_{xym} = P(R_m = 0 \mid Z_m = 1, x, y) \tag{6}$$

$$\phi'_{xym} = P(R_m = 1 \mid Z_m = 0, x, y), \tag{7}$$

for each m, is sufficient to uniquely determine the matrix A_{xy}. This amounts to $2k$ parameters $\Phi_{xy} = (\phi_{xy1}, \ldots, \phi_{xyk})$ and $\Phi'_{xy} = (\phi'_{xy1}, \ldots, \phi'_{xyk})$ in total (as opposed to 2^{2k} needed for

A_{xy}). The IF pattern can be naturally combined with the no false positives assumption in Eq. 2, in which case $\Phi'_{xy} = (0, \dots, 0)$ for each x, y. Finally, a further assumption that may be considered is that of parameter sharing (PS, for short), which is in this context stated as:

$$P(r_m \mid z_m, x, y) = P(r_\ell \mid z_\ell, x, y) = \phi_{xy} \text{ whenever } z_m = z_\ell, r_m = r_\ell \text{ and } \forall x, y, \tag{8}$$

meaning that the entries of Φ_{xy} (or Φ'_{xy}) are all equal. Under both PS (Eq. 8) and NFP (Eq. 2), the matrix A_{xy} is determined by the single parameter ϕ_{xy}, and given by

$$(A_{xy})_{ij} = P(r^{(i)} \mid z^{(j)}, x, y) = \begin{cases} 0 & \text{if } \exists\, m \text{ s.t. } z_m^{(j)} < r_m^{(i)} \\ \phi_{xy}^{\|z^{(j)} - r^{(i)}\|_0} (1 - \phi_{xy})^{\|r^{(i)}\|_0} & \text{otherwise,} \end{cases} \tag{9}$$

where by $\| \cdot \|_p$ we denote the ℓ_p-norm.

2.2.1.2. Zero-Inflation. Another interesting fidelity pattern we consider is zero-inflation (ZINF, for short). When considering this pattern, we assume that measurement errors are such that

$$P(R = 0 \mid Z = z, X = x, Y = y) = \phi_{zxy} \tag{10}$$

$$P(R = z \mid Z = z, X = x, Y = y) = 1 - \phi_{zxy}. \tag{11}$$

In words, for any value $Z = z$ (and given $X = x, Y = y$), we record the correct value with probability $1 - \phi_{zxy}$, while all the values are recorded as 0 with probability ϕ_{zxy}. This fidelity pattern requires the specification of $2^k - 1$ parameters $\Phi_{xy} = (\phi_{z^{(1)}xy}, \dots, \phi_{z^{(2^k-1)}xy})$ to determine A_{xy} uniquely, and also implies the NFP assumption (Eq. 2). Once again, we can use a parameter sharing assumption, which states that

$$\phi_{zxy} = \phi_{z'xy} = \phi_{xy}\ \forall z, z' \neq 0. \tag{12}$$

With parameter sharing, the matrix A_{xy} is uniquely determined by a single parameter ϕ_{xy}.

Arguably, other interesting simplifications of fidelity patterns may exist, and the application context may help guide the analyst to consider appropriate fidelity patterns. We remark that the developments in the remainder of the paper consider the general case of arbitrary A_{xy} fidely patterns, and can thus be adapted to such alternative patterns.

2.3. *A Sensitivity Approach for Measurement Error.*

We are now ready to introduce the sensitivity analysis based on the ϕ_{xy} parameters. Denote by Φ the vector $(\Phi_{x_0y_0}, \Phi_{x_1y_0}, \Phi_{x_0y_1}, \Phi_{x_1y_1})$, and let the $\text{ATE}_{x_0,x_1}(y; \Phi)$ be the causal effect estimate assuming $\{A_{xy} : x, y \in \{0, 1\}\}$ computed based on Φ_{xy} parameters. Consider the following definition:

DEFINITION 3 (ϕ-Value). *Let $\mathcal{M}$ be a structural causal model compatible with the causal diagram in Fig. 1a. Let $ATE_{x_0,x_1}(y; 0)$ denote the average treatment effect of X on Y assuming the vector $\Phi = (0, 0, 0, 0)$ (no measurement error). Assume without loss of generality that $ATE_{x_0,x_1}(y; 0) \geq 0$. Then, a ϕ-value is any fidelity pattern Φ such that*

$$ATE_{x_0,x_1}(y; \Phi) < 0. \tag{13}$$

The intuition behind the notion of ϕ-values is simple. If under no measurement error ($\Phi = 0$) the effect is estimated to be positive, then a ϕ-value is any fidelity pattern under which the effect estimate would change sign. In practice, we are often interested in minimal ϕ-values with respect to some metric. A natural choice may be the ℓ_1-norm, which would implicitly encourage sparse ϕ-values that may be more easily interpreted. Furthermore, it is helpful to distinguish between agnostic, x-specific, y-specific, and xy-specific ϕ-values according to Def. 2.

2.4. *Estimators and Guarantees.* In this section, we discuss two approaches for estimation of causal effects under imperfect observations, $\text{ATE}_{x_0,x_1}(y;\Phi)$, in turn allowing us to infer ϕ-values. First, we discuss a non-parametric approach that makes no distributional assumptions over $P(Z,X,Y)$. Then, we discuss a parametric approach using expectation-maximization (EM) (Dempster et al., 1977) for the case when the distributions $P(Z), P(X \mid Z), P(Y \mid X, Z)$ come from specific exponential families.

2.4.1. *Non-parametric Estimation.* The non-parametric estimator we propose is motivated by the recovery conditions obtained in Thm. 1. In particular, a sample-level bin-counting estimator of the population-level identification expression in Supplement A is used, with the term $P(Y_x = 1)$ estimated by

(14)
$$\widehat{Y}_x(\Phi) := \xi \left[A^{-1}_{xy_1}(\Phi)\widehat{p}_{r|xy_1} \oslash \Big(\sum_{y'} \widehat{p}_{y'|x} A^{-1}_{xy'}(\Phi)\widehat{p}_{r|x_1y'} \Big) \widehat{p}_{y_1|x} \right] \cdot \left(\sum_{x',y'} A^{-1}_{x'y'}(\Phi)\widehat{p}_{r|x'y'} \right)$$

where $\widehat{p}_{r|xy} := \frac{1}{n(x,y)} \sum_{s=1}^{n} \mathbb{1}(X^s = x, Y^s = y)\big(\mathbb{1}(R^s = r^{(0)}), \ldots, \mathbb{1}(R^s = r^{(2^k-1)})\big)$ is the bin-counting estimator of the population-level $p_{r|xy}$. Here, $n(x,y) := \sum_{s=1}^{n} \mathbb{1}(X^s = x, Y^s = y)$ is the number of samples with $X = x, Y = y$, while $\widehat{p}_{y|x}$ is shorthand for the empirical estimate $\hat{P}(x,y) = \frac{n(x,y)}{n}$. The $\xi(x) = (x \vee 0) \wedge 1$ trims each vector entry to the unit interval. The trim function is used since the input to ξ in Eq. 14, which attempts to estimate the probability $P(y \mid x, z) \in [0,1]$, does not necessarily lie in the unit interval (a more detailed discussion can be found in the next section). The estimator of $\text{ATE}_{x_0,x_1}(y;\Phi)$ is then given by:

(15)
$$\widehat{\tau} := \widehat{Y}_{x_1} - \widehat{Y}_{x_0}.$$

2.4.2. *Well-posedness of the Inverse Problem.* The above framing can be viewed as an instance of an inverse problem. Given observations R, we are interested in inferring effects based on the true underlying confounders Z. A common concern in the literature on inverse problems (Groetsch and Groetsch, 1993; Calvetti and Somersalo, 2018) is *well-posedness*. In particular, for computing the $\widehat{\tau}$ estimator in Eq. 15, we are implicitly computing conditional probabilities $P(z \mid x, y)$ for all values of z, x, y, based on the transformation $p_{z|xy} = A^{-1}_{xy} p_{r|xy}$ following from Eq. 4. Depending on the empirical values that are observed, say $\widehat{p}_{r|xy}$, the resulting estimator of $p_{z|xy}$, labeled $\widehat{p}_{z|xy}$, need not lie on the unit simplex, since it may contain negative entries (intuitively, we may specify the measurement error as too large, not compatible with the observed data). In such cases, for a specific choice of the data fidelity Φ, we say the inverse problem is ill-posed. To handle these cases, we propose to replace the inverse problem with a similar problem that is well-posed. In particular, suppose that $\widehat{p}_{z|xy}$ has negative entries for some initial Φ^{init}. Clearly, for $\Phi = 0$, $\widehat{p}_{z|xy} = \widehat{p}_{r|xy}$ constitutes a valid problem (vector $\widehat{p}_{r|xy}$ lies on the unit simplex by definition). Therefore, let $\eta(\widehat{p}_{z|xy}) = \|\widehat{p}_{z|xy} \wedge 0\|_1$ denote the sum of absolute values of the negative entries of $\widehat{p}_{z|xy}$. When facing an ill-posed problem, gradient descent steps are performed, minimizing $\eta(\widehat{p}_{z|xy})$, via updates

(16)
$$\Phi^{(T+1)}_{xy} = \Phi^{(T)}_{xy} + \alpha \cdot (\nabla_{\Phi_{xy}} \eta(\widehat{p}_{z|xy}) \wedge 0),$$

where the minimum operator $\wedge$ in Eq. 16 ensures that each gradient step is towards the origin. In practice, the step size α is chosen adaptively using the Armijo-Goldstein criterion (Armijo, 1966). When the well-posed setting of fidelity parameters Φ^{wp}_{xy} is found, it needs to be reported, and depending on the type of ϕ-value (agnostic, x, y or x,y-specific) also be used for other values of X, Y, i.e., setting $\Phi_{x'y'} = \Phi^{\text{wp}}_{xy}$ for x', y' possibly different from x, y. For instance, for agnostic ϕ-values, where the fidelity should not depend on X, Y, we set each $\Phi_{x'y'}$ to Φ^{wp}_{xy} (and analogously for x-, y-, or xy-specific cases).

2.4.3. *Convergence of the Non-Parametric Estimator.* We now provide a convergence rate bound for the non-parametric estimator:

THEOREM 2 (Convergence rate of $\widehat{\tau}$). *Let n be the number of samples (X^s, Y^s, R^s), k the dimensionality of the space of confounders Z, and matrices $\{A_{xy}\}_{x,y\in\{0,1\}}$ the true fidelity pattern. Suppose that the following assumptions hold:*

(i) $\min_{z,x} P(X = x \mid Z = z) > c_1 > 0$,
(ii) $\min_{x,y} P(x, y) > c_2 > 0$,
(iii) *each A_{xy} is invertible, with $\lambda_{\min}(A_{xy})$ denoting the smallest absolute eigenvalue of A_{xy}, and $\xi(A_{xy}^{-1})$ the largest absolute entry $\sup_{i,j} |(A_{xy}^{-1})_{ij}|$ of A_{xy}^{-1}.*

The estimator $\widehat{\tau}(\Phi)$ from Eq. 15 is consistent for $\tau(\Phi)$, and for any $\delta \in (0,1)$ there exist universal constants M_1, M_2 such that

$$(17)\qquad \big(\tau(\Phi) - \hat{\tau}(\Phi)\big)^2 \leq \frac{1}{n\delta}\left\{M_1 \cdot 2^{5k} \max_{x,y} \xi(A_{xy}^{-1})^2 + M_2 \cdot \frac{2^{3k}}{\min_{x,y} \lambda_{\min}(A_{xy})^2}\right\}$$

with probability at least $1 - \delta$ for any $n \geq \frac{4(2^k+1)}{\delta c_2^2}$.

The theorem's proof is given in Supplement B. The theorem formally demonstrates several points that one may expect from the $\hat{\tau}(\Phi)$ estimator. Firstly, the estimator's error is exponential in the number of dimensions k. Secondly, the error also grows as (i) $\lambda_{\min}(A_{xy})$ approaches zero, that as A_{xy} matrices come closer to a singular matrix; (ii) as the maximal entry of A_{xy}^{-1} grows large. For instance, under the assumptions NFP (Eq. 2), IF (Eq. 5), and PS (Eq. 8), we have that $\lambda_{\min}(A_{xy})^{-1} = \xi(A_{xy}^{-1}) = (1 - \phi_{xy})^{-k}$. In this case, the first term on the RHS of Eq. 17 dominates, and the estimator's error grows exponentially as ϕ_{xy} (probability of omitting a confounder) grows larger.

2.4.4. *Exponential Families – Parametric Estimation.* An alternative approach for inference is to assume a parametric family form for the distributions $P(Z), P(X \mid Z)$, and $P(Y \mid X, Z)$. Here, for $P(Z)$ we suggest an exponential family model with first-order interactions, given through

$$(18)\qquad P(z) = \frac{1}{A(\Sigma)} \exp(z^T \Sigma z),$$

where Σ is a symmetric matrix in $\mathbb{R}^{k\times k}$, and $A(\cdot)$ is the cumulant function. The key feature of this modeling choice is that first-order interactions are possible, that is, $P(Z_m = z_m, Z_\ell = z_\ell) \neq P(Z_m = z_m)P(Z_\ell = z_\ell)$ in general. This is important for the application in question since certain groups of comorbidities may be correlated (for instance, impaired kidney function and diabetes). We further assume that X and Y follow logistic models with an intercept, namely:

$$(19)\qquad P(X = 1 \mid Z = z) = \text{expit}(\lambda^T(1, z)),$$

$$(20)\qquad P(Y = 1 \mid Z = z, X = x) = \text{expit}(\mu^T(1, z) + \beta x),$$

where $\text{expit}(a) = \frac{\exp(a)}{1+\exp(a)}$. Let us denote the model parameters by $\theta = (\Sigma, \lambda, \mu, \beta)$. To infer these parameters from finite sample data, we use the expectation-maximization (EM) algorithm (Dempster et al., 1977). We write the Q-function that is optimized using the EM algorithm:

$$(21)\qquad Q(\theta' \mid \theta) = \sum_{r,x,y} P_{\theta^*}(r, x, y) \sum_z P_\theta(z \mid r, x, y) \log P_{\theta'}(z, r, x, y).$$

The sample version of Q, denoted by Q_n, is obtained by replacing the true distribution $P_{\theta^*}(r,x,y)$ by the empirical distribution $\hat{P}(r,x,y)$. Importantly, the joint likelihood factorizes as follows:

$$P_\theta(z,r,x,y) = P_\Sigma(z)P_\lambda(x \mid z)P_{\mu,\beta}(y \mid x,z)P_\Phi(r \mid x,y,z). \tag{22}$$

This allows us to re-write $Q(\theta' \mid \theta)$ as

$$\sum_{r,x,y} P_{\theta^*}(r,x,y)\sum_{z} P_\theta(z \mid r,x,y)\Big[\log P_{\Sigma'}(z) + \log P_{\lambda'}(x \mid z) + \log P_{\mu',\beta'}(y \mid x,z) + \log P_\Phi(r \mid x,y,z)\Big], \tag{23}$$

showing we can optimize for Σ', λ', and (μ', β') separately. We do so using Monte Carlo EM.

2.4.4.1. Inferring θ from Monte Carlo Samples. Let $(X^i, Y^i, Z^i)_{i=1}^M$ be a Monte Carlo sample of (X,Y,Z) based on current parameters θ. The Σ parameter can be inferred from the second moments $\mathbb{E}[Z_m Z_\ell]$ of the exponential family in Eq. 18. Second moments are easily estimated from the MC sample of size n_M:

$$\hat{\mathbb{E}}[Z_m Z_\ell] = \frac{1}{n_M}\sum_{s=1}^{n_M} Z_m^s Z_\ell^s. \tag{24}$$

Crucially, for any set of values of second moments, there exists a solution Σ compatible with these moments, due to a known result on the expressiveness of exponential families:

Proposition 3 (Exponential Family Representation (Wainwright et al., 2008)). *Let $Z \sim F_Z$ be distributed over $\{0,1\}^k$, and let $\mu = \mathbb{E}[ZZ^T]$ be the second moments of F_Z. Then there exists Σ such that for the exponential family in Eq. 18 we have $\mathbb{E}_\Sigma[ZZ^T] = \mu$.*

In words, all mean parameters (in this case second moments of Z) that are realizable by any distribution can be realized by a member of the exponential family in Eq. 18. The maximum likelihood estimator (MLE) of Σ maximizes the log-likelihood with empirical moments,

$$\widehat{\Sigma} = \arg\max_{\Sigma} \sum_{m,\ell} \widehat{\mathbb{E}}[Z_m Z_\ell] \cdot \Sigma_{m\ell} - \log(A(\Sigma)). \tag{25}$$

The derivative of the objective with respect to $\Sigma_{m\ell}$ is given by

$$\frac{\partial \log L(\Sigma)}{\partial \Sigma_{m\ell}} = \widehat{\mathbb{E}}[Z_m Z_\ell] - \frac{1}{A(\Sigma)}\frac{\partial A(\Sigma)}{\partial \Sigma_{m\ell}}, \tag{26}$$

and the second derivative is just the Hessian matrix of the distribution, $H(\Sigma)$. Thus, as usual with exponential families, we can find the MLE $\widehat{\Sigma}$ via the Newton-Raphson algorithm. Similarly, for the λ parameter, we regress X onto Z using the same Monte-Carlo sample, and for (μ, β) we regress Y onto Z, X.

2.4.4.2. Monte-Carlo EM Formulation. Putting everything together, we can describe the steps of our Monte-Carlo EM approach:

(0) Get an initial estimate of $\widehat{\lambda}^{(0)}, \widehat{\mu}^{(0)}, \widehat{\beta}^{(0)}$ by performing a logistic regression of X onto R, and Y onto R and X. For the initial estimate of $\widehat{\Sigma}^{(0)}$, suppose that $Z = R$, and infer $\widehat{\Sigma}^{(0)}$ from the empirical second moments $\hat{\mathbb{E}}[R_m R_\ell]$.

(1) At iteration $T = t$, for each sample $R^s = r, X^s = x, Y^s = y$, compute the distribution

$$P_{\hat{\theta},\Phi}(z \mid r,x,y) \tag{27}$$

using the parameters $\hat{\theta}^{(t)} = (\widehat{\Sigma}^{(t)}, \widehat{\lambda}^{(t)}, \widehat{\mu}^{(t)}, \widehat{\beta}^{(t)})$, and Φ. Draw n_{mc} Monte Carlo samples from this distribution, labeled $Z^{s,1,(t)}, \ldots, Z^{s,n_{mc},(t)}$. Add samples $(Z^{s,b,(t)}, X^s, Y^s)_{b=1}^{n_{mc}}$ to the complete data Monte Carlo sample at iteration $T = t$, labeled $\mathcal{D}^{(t)}$.

(2) Using $Z^{(t)}$, estimate the Monte-Carlo empirical second moments of Z, and infer $\widehat{\Sigma}^{(t+1)}$ based on them.
(3) Using $\mathcal{D}^{(t)}$ estimate $\widehat{\lambda}^{(t+1)}, \widehat{\mu}^{(t+1)}, \widehat{\beta}^{(t+1)}$ by regressing X onto $Z^{(t)}$ and Y onto $Z^{(t)}, X$.
(4) Increase $t \rightarrow t+1$ and repeat steps (1) -(3) until convergence.

After T iterations, the estimate $\widehat{\beta}^{(T)}$ can be used as an estimator of the true effect β. $\widehat{\beta}^{(T)}$ can also be used in combination with $\widehat{\Sigma}, \widehat{\lambda}^{(T)}, \widehat{\mu}^{(T)}$ to compute the ATE given by $\sum_z [P_{\hat{\theta}^{(T)}}(y \mid x_1, z) - P_{\hat{\theta}^{(T)}}(y \mid x_0, z)] P_{\hat{\theta}^{(T)}}(z)$. Uncertainty quantification for the parameter $\widehat{\beta}^{(T)}$ can be performed using Louis' method (Louis, 1984).

2.4.4.3. On Providing Theoretical Guarantees. The work of Balakrishnan et al. (2017) provides a useful set of general conditions for demonstrating the convergence of EM estimators. However, when attempting to apply this approach to our setting, it turns out that the conditions proposed by Balakrishnan et al. (2017) are not satisfied. Therefore, this theoretical framework does not apply in our case. This reflects a limitation of the specific sufficient (but not necessary) conditions provided in Balakrishnan et al. (2017), rather than a general impossibility of obtaining guarantees for our sensitivity model. A detailed discussion of this is provided in Supplement C, together with the relevant empirical results.

2.4.5. *Continuous-Discrete Mixed Data.* The methodology proposed so far assumed binary confounders subject to measurement error. A useful extension of this setting is to consider additional, correctly observed continuous confounders. Let W taking values in $\mathbb{R}^d$ denote the continuous confounders assumed to have no measurement error. In this case, we propose an exponential family of the form:

$$P(z, w) = \frac{1}{A(\Sigma, \Lambda, \Omega, b)} \exp(z^T \Sigma z + w^T \Lambda z - \frac{1}{2} w^T \Omega w + b^T w), \tag{28}$$

while assuming that the measurement error pattern satisfies $P(r \mid z, x, y, w) = P(r \mid z, x, y)$, meaning that the pattern does not depend on the continuous variables W. The parameterization in Eq. 28 allows interactions between binary confounders Z and continuous confounders W, governed by the Λ parameter, while the b parameter allows the mean of W to vary freely. Distributional forms for X, Y in Eqs. 19-20 are also extended to include W. In Supplement D, we derive tractable expressions for the gradient and Hessian of this exponential family, allowing one to use Newton-Raphson steps to optimize likelihood in a similar fashion as for the exponential family in Eq. 18.

2.4.6. *Computational Complexity.* Another key consideration for the proposed inference approaches is their computational complexity, which we now briefly discuss. For constructing the non-parametric estimator $\widehat{\tau}$, we first need to assign each sample (R^s, X^s, Y^s) to its respective bin, taking $\mathcal{O}(n)$ steps. Constructing the vectors $\widehat{p}_{r|x,y}$ depends exponentially on the dimension k, and requires $\mathcal{O}(2^k)$ steps. Then, the most expensive step is the computation of $\widehat{p}_{z|x,y}$ given by $A_{xy}^{-1} \widehat{p}_{r|x,y}$. The matrix A_{xy} is of size $2^k \times 2^k$ and can thus be constructed in $\mathcal{O}(2^k \cdot 2^k)$ steps, and taking its inverse requires $\mathcal{O}(2^{3k})$ steps. However, for independent fidelity and zero-inflation patterns the matrix A_{xy}^{-1} can be computed more efficiently:

LEMMA 1 (A_{xy}^{-1} under Different Fidelity Patterns). *Let A_{xy} be the fidelity pattern. Consider the following two cases:*

(i) *Under assumptions of no false positives (Eq. 2), independent fidelity (Eq. 5), and parameter sharing (Eq. 8), the entries of the matrix A_{xy}^{-1} can be computed as*

$$(A_{xy}^{-1})_{ij} = (-1)^{\|r^{(j)} - z^{(i)}\|_0} \phi_{xy}^{\|r^{(j)} - z^{(i)}\|_0} (1 - \phi_{xy})^{-\|r^{(j)}\|_0} \mathbb{1}(r^{(j)} \geq z^{(i)}). \tag{29}$$

(ii) *Under assumption of zero-inflation (Eq. 10), the entries of the matrix A_{xy}^{-1} can be computed as*

$$(A_{xy}^{-1})_{ii} = \frac{1}{1-\phi_{r(i)}} \; \forall i, \quad (A_{xy}^{-1})_{1i} = \frac{-\phi_{r(i)}}{1-\phi_{r(i)}} \; \forall i > 1, \tag{30}$$

and $(A_{xy}^{-1})_{ij} = 0$ *otherwise.*

The lemma allows us to compute the inverse matrix in $\mathcal{O}(k \cdot 2^{2k})$ steps for the case of independent fidelity, and $\mathcal{O}(k \cdot 2^k)$ for the case of zero-inflation. This offers a substantial speed-up compared to $\mathcal{O}(2^{3k})$ steps required for the general case. Therefore, the estimator $\widehat{\tau}$ can be computed with an overall computational complexity of $\mathcal{O}(n + k \cdot 2^{2k})$ under such simplifying assumptions, whereas $\mathcal{O}(n + 2^{3k})$ steps are needed for the most general setting.

We now turn to analyzing the complexity of the two-step parametric approach. Let n_{mc} be the number of Monte Carlo samples for each observed sample s. In the expectation step, for each of the n samples we draw n_{mc} samples. Computing the conditional probabilities $P(z \mid r, x, y)$ for a fixed choice of $(R^s, X^s, Y^s) = (r, x, y)$ requires $\mathcal{O}(2^k k^2)$ steps, yielding $\mathcal{O}(n n_{mc} 2^k k^2)$ for the E-step.

In the M-step, to infer the MLE of Σ, we need to compute $\widehat{\mathbb{E}}[ZZ^T]$, which requires $\mathcal{O}(k^2 n n_{mc})$ steps. After this, a Newton-Raphson procedure is applied, and the number of iterations to get enough precision scales approximately with $\log n$. Each iteration requires the computation of a gradient and the inverted Hessian, with the inverse Hessian computation being more costly at $\mathcal{O}(k^6)$, yielding a total $\mathcal{O}(k^2 n n_{mc} + k^6 \log n)$. For inferring (λ, μ, β) we fit logistic models. In a single step of the Newton-Raphson method, we compute the Hessian ($n n_{mc} k^2$ steps), the gradient ($n n_{mc} k$), and the update step (k^3). If the Newton-Raphson method requires $\mathcal{O}(1)$ iterations to converge for a logistic model (Chambers and Hastie, 2017), we arrive at $\mathcal{O}(n_{mc} n k^2 + k^3)$ for this part, meaning that the M-step is dominated by Σ inference, with complexity $\mathcal{O}(k^2 n n_{mc} + k^6 \log n)$.

Assuming that $n_{mc} = \mathcal{O}(1)$ and that the number of iterations required for the EM algorithm to converge to a $\frac{1}{\sqrt{n}}$ neighborhood of the true parameter scales as $\mathcal{O}(\log n)$ (Balakrishnan et al., 2017), the overall complexity is $\mathcal{O}(2^k k^2 n \log n + k^6 \log^2 n)$. After analyzing the statistical and computational properties of the two methods, we apply them to the original problem.

3. Results.

We next infer ϕ-values on the motivating application. In particular, we use the MIMIC-IV v2.2 (Johnson et al., 2021) dataset from the Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts. Data loading is performed using the `ricu` R-package (Bennett et al., 2023).

For the patient cohort, we focus on all adult patients admitted to the hospital between 2008 and 2019, who in the electronic health record of the hospital have a body-mass index (BMI) in the range $[18.5, 35)\,\mathrm{kg/m^2}$. Underweight (BMI $< 18.5\mathrm{kg/m^2}$) and morbidly obese (BMI $\geq 35\mathrm{kg/m^2}$) patients are excluded from the analysis, since these groups are known to have poor outcomes (Hainer and Aldhoon-Hainerová, 2013). Therefore, we focus on the clinically most relevant aspect of the obesity paradox, the comparison between normal and overweight/class I obese patients. The resulting cohort consists of $n = 99,417$ patients.

The outcome Y is in-hospital mortality, determined from discharge records. Treatment X is defined using body mass index (BMI), calculated from height and weight entries recorded in the hospital electronic health record (x_0 for BMI < 25, x_1 for BMI ≥ 25). Confounders Z are comorbidities in the Elixhauser Index (Elixhauser et al., 1998), extracted from International Classification of Diseases (ICD-9 and ICD-10) diagnosis codes documented at the time of hospital discharge from the medical record. All variables are obtained from structured EHR data recorded at BIDMC, with no information from outside hospitals. The sensitivity analysis carried out aims to examine the effect of possible errors in the data collection process (i.e., comorbidities not mentioned in the medical record, or not coded into an ICD code) and how these measurement errors may affect the observed obesity paradox in the data.

Next, we regress Y onto X and Z in a logistic model (i.e., assuming there is no measurement error). To reduce the dimensionality of Z, we consider all comorbidities which have a statistically significant association with increased mortality (at 5% significance level). This yields the following 14 comorbidities $Z_1, \dots, Z_{14}$ ordered in decreasing effect size: Fluid and Electrolyte Disorders, Metastatic Cancer, Cardiac Arrhythmia, Liver Disease, Coagulopathy, Other Neurological Disorders, Paralysis, Congestive Heart Failure, Lymphoma, Peripheral Vascular Disease, Renal Failure, Pulmonary Hypertension, Chronic Pulmonary Disease, Solid Tumor (excluding metastatic cancer).

Multiple previous studies (Hainer and Aldhoon-Hainerová, 2013; Hutagalung et al., 2011) find that the effect of increased BMI on mortality Y is protective (reduces mortality), written in our notation as

$$\text{ATE}_{x_0,x_1}(y; \Phi = 0) < 0. \tag{31}$$

This finding is also replicated in our data. However, the effect estimate may be affected by measurement error. To explore this, we search for ϕ-values – minimal patterns of (informative) measurement error – that explain away the counterintuitive protective effect of obesity.

In this section, we assume there are no false positives (Eq. 2), which is justifiable in our application since comorbidities that are not present are unlikely to be recorded as present. We analyze the independent fidelity pattern, defined by (recalling Eq. 5)

$$P(R \mid Z, X, Y) = \Pi_{m=1}^{k} P(R_m \mid Z_m, X, Y),$$

and the zero-inflation pattern, defined through (recalling Eq. 10)

$$P(R = 0 \mid Z = z, X, Y) = \phi_z, P(R = z \mid Z = z, X, Y) = 1 - \phi_z \ \forall z.$$

In the sequel, sensitivity analyses for measurement error are performed on real data for both non-parametric and parametric approaches, while Supplement F contains experiments for semi-synthetic data with a known ground truth, verifying the validity of the implementation.

3.1. *Non-parametric Inference of ϕ-values.* We begin with non-parametric inference of ϕ-values. Since the number of samples required for this approach scales exponentially with dimension k, we reduce the dimension of the set of confounders by constructing a clustering over the 14 comorbidities into $k = 6$ groups as follows: Cardiovascular (Cardiac Arrhythmia, Congestive Heart Failure, Peripheral Vascular Disease, Pulmonary Hypertension), Renal-Metabolic (Renal Failure, Fluid and Electrolyte Disorders), Pulmonary (Chronic Pulmonary Disease), Neurological (Paralysis, Other Neurological Disorders), Malignancy (Solid Tumor excluding metastatic cancer, Metastatic Cancer, Lymphoma), Liver-Coagulation (Liver Disease, Coagulopathy). These groups correspond to common organ system-based considerations found in comorbidity indices and intensive care illness severity scores (Charlson et al., 1987; Le Gall et al., 1993; Vincent et al., 1996). In each group $g \in \{1, \dots, 6\}$, we construct a binary variable $\tilde{Z}_g = \max_{m \in I_g} Z_m$ (where I_g is the index set of group g), indicating whether any of the comorbidities in the group is recorded as active for the patient. Binary indicators $\tilde{Z}_1, \dots, \tilde{Z}_6$ of each cluster are used for the non-parametric analysis.

When investigating independent fidelity (Eq. 5) and zero-inflation (Eq. 10) patterns, for simplicity we assume parameter sharing (Eqs. 8, 12). We search for agnostic, x-specific, and y-specific ϕ-values following Def. 2. The results are summarized in Fig. 2. Fig. 2(a) shows that agnostic measurement error is unlikely to explain away the obesity paradox, since over a range of values of ϕ, no major changes in the estimated ATE are found. We next search

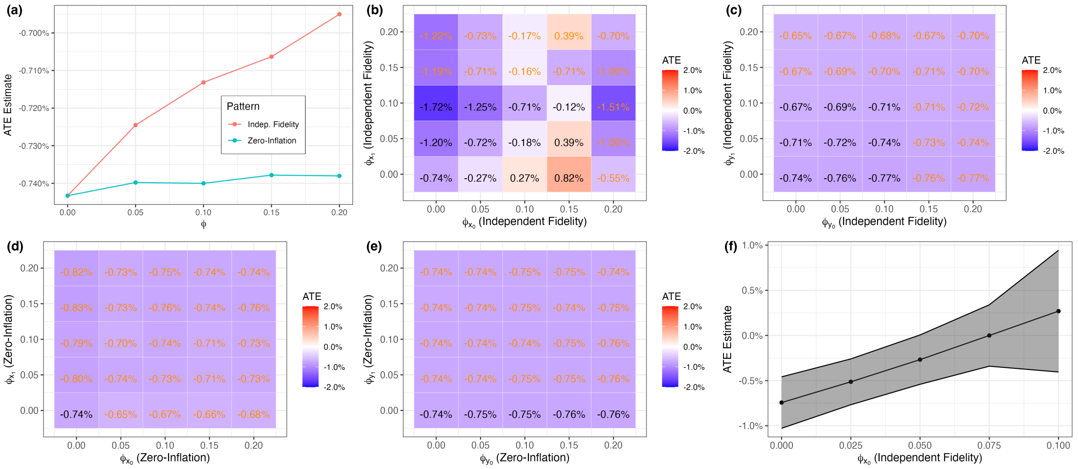


Fig 2: Non-parametric inference of ϕ-values: (a) agnostic; (b) x-specific independent fidelity; (c) y-specific independent fidelity; (d) x-specific zero-inflation; (e) y-specific zero-inflation; (f) x-specific independent fidelity with uncertainty quantification for increasing ϕ_{x_0}.

for x-specific and y-specific ϕ-values, summarized in Figs. 2(b-e). In the figures, orange text color indicates instances of Φ which were found to be incompatible with the observed data, meaning that the inverse problem was ill-posed, and the effective values of Φ had to be adjusted (see Sec. 2.4.2). Importantly, the y-specific search does not point to a clear trend, neither for independent fidelity nor zero-inflation patterns. The x-specific grid search indicates $\text{ATE}_{x_0,x_1}(y;\Phi)$ values approaching 0 for increasing ϕ_{x_0} and fixed $\phi_{x_1} = 0$. For independent fidelity, the pattern $(\phi_{x_0}, \phi_{x_1}) = (0.1, 0)$ changes the sign of the point estimate of the ATE (see Fig. 2(b)). To investigate further, the x-specific search along this direction is repeated over 500 bootstrap samples of the data, to analyze the effect of finite sample uncertainty. The results are shown in Fig. 2(f), indicating that for $\phi_{x_0} = 0.05$ the 95% confidence interval for $\text{ATE}_{x_0,x_1}(y;\Phi)$ includes 0, meaning that the null hypothesis $H_0 : \text{ATE}_{x_0,x_1}(y;\Phi) \geq 0$ cannot be rejected at 2.5% significance level under this level of measurement error.

3.2. *Exponential Family Inference of ϕ-values.* Inference of ϕ-values based on the parametric approach is performed next (see Sec. 2.4.4). In particular, we assume that the data is sampled from the model in Eqs. 18-20, and the EM procedure from Sec. 2.4.4.2 is used. Since the number of parameters scales only quadratically in the dimension of the confounders Z, we use all of $Z_1, \ldots, Z_{14}$ for this experiment. The results of searches for agnostic, x-specific, and y-specific ϕ-values are summarized in Fig. 3. Fig. 3(a) shows that agnostic independent fidelity ϕ-values point to a trend, but are still unlikely to explain away the effect under study. The agnostic zero-inflation pattern does not point to a clear trend. Figs. 3(c, e) show that y-specific ϕ-values also do not point to a specific trend. Finally, for x-specific ϕ-values (Figs. 3(b, d)), ATE estimates approach 0 in the direction of increasing ϕ_{x_0} and $\phi_{x_1} = 0$, with a faster trend for the independent fidelity pattern. Furthermore, our approach for detecting ill-posedness (Sec. 2.4.2) finds $\phi_{x_0} > 0.05$ values for the zero-inflation pattern are unlikely to be compatible with the observed data. Therefore, we perform a more detailed search along the x-specific direction for the independent fidelity pattern, with confidence intervals obtained using the method of Louis (1984). Fig. 3(f) shows that for the x-specific independent fidelity ϕ-value $(\phi_{x_0} = 0.05, \phi_{x_1} = 0)$ the null hypothesis $H_0 : \text{ATE}_{x_0,x_1}(y;\Phi) \geq 0$ would not be rejected at 2.5% significance level, implying that this level of measurement error may explain away the causal effect under study.

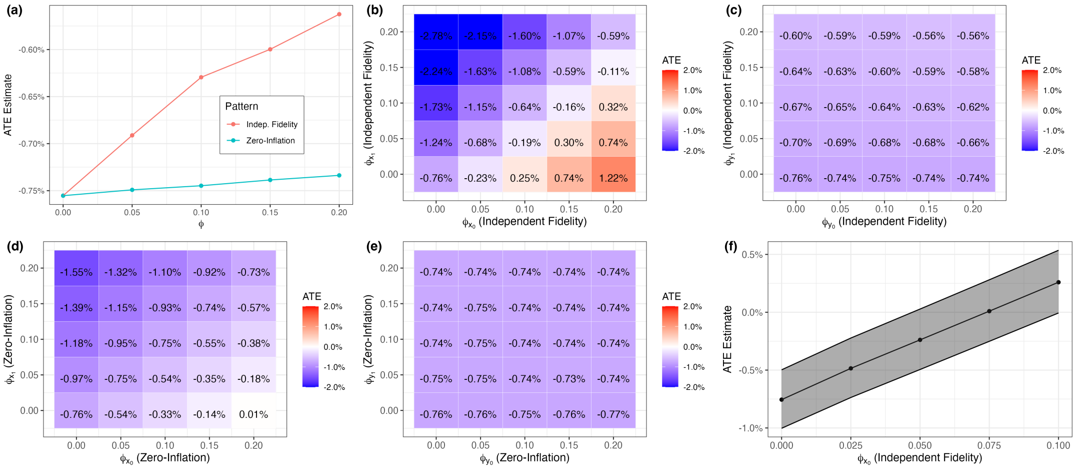


Fig 3: Parametric inference of ϕ-values: (a) agnostic; (b) x-specific independent fidelity; (c) y-specific independent fidelity; (d) x-specific zero-inflation; (e) y-specific zero-inflation; (f) x-specific independent fidelity with uncertainty quantification for increasing ϕ_{x_0}.

3.3. *Extending with Continuous Confounders – Effect of Age.* After determining possible measurement errors that could explain away the obesity paradox when considering comorbidities as confounders, we next consider a continuous confounding variable that may be important in our context – age. Age is known to monotonically increase the risk of death and a number of comorbidities (Charlson et al., 1987; Marengoni et al., 2011). The impact of age on the obesity paradox, however, has not been clearly established so far (Bosello and Vanzo, 2021). To investigate how age affects our analysis, we first plot the probability $P(X = 1 \mid W = w)$ of overweight/obesity across age groups $W = w$, shown in Fig. 4(a). As we can see, the probability of overweight/obesity follows an inverted U-shaped relationship with age. Thus, age may be an interesting confounder due to an established relationship with Z, X, Y.

Following the proposal in Sec. 2.4.5, we use Monte Carlo EM to fit an exponential family model for Z, W as in Eq. 28. The age variable is normalized to have mean 0, and variance 1, and we label the normalized variable $\tilde{W}$. Additionally, we also include $\tilde{W}^2$ in the model for the X variable, due to the U-shaped relationship of age and BMI. As before, a search for ϕ-values is performed in the direction of increasing ϕ_{x_0} and we restrict our attention to the independent fidelity pattern (IF). The results are shown Fig. 4(b), and we observe different effect estimates compared to not including age. However, we again find the same ϕ-value for the x-specific independent fidelity, namely $(\phi_{x_0} = 0.05, \phi_{x_1} = 0)$, for which the null hypothesis $H_0 : \text{ATE}_{x_0,x_1}(y; \Phi) \geq 0$ would not be rejected at 2.5% significance level.

4. Conclusion.

In this manuscript, we introduced a type of sensitivity analysis for average treatment effects in cases of measurement error on binary confounders. In particular, the sensitivity analysis answers the question "If the binary confounders are recorded imperfectly with a probability ϕ, how large would ϕ have to be to explain away the causal effect under study, that is, reverse the sign of the causal effect estimate?". Our results indicate that answering this question is possible in surprising generality (Thm. 1), including cases in which confounder values $Z_m = 1$ are recorded as $R_m = 0$ (false negatives), and cases where values $Z_m = 0$ are recorded as $R_m = 1$ (false positives). Furthermore, the pattern of measurement error is allowed to depend both on the treatment X and the outcome Y (Def. 2). We developed two formal approaches for performing the sensitivity analysis: a non-parametric approach based

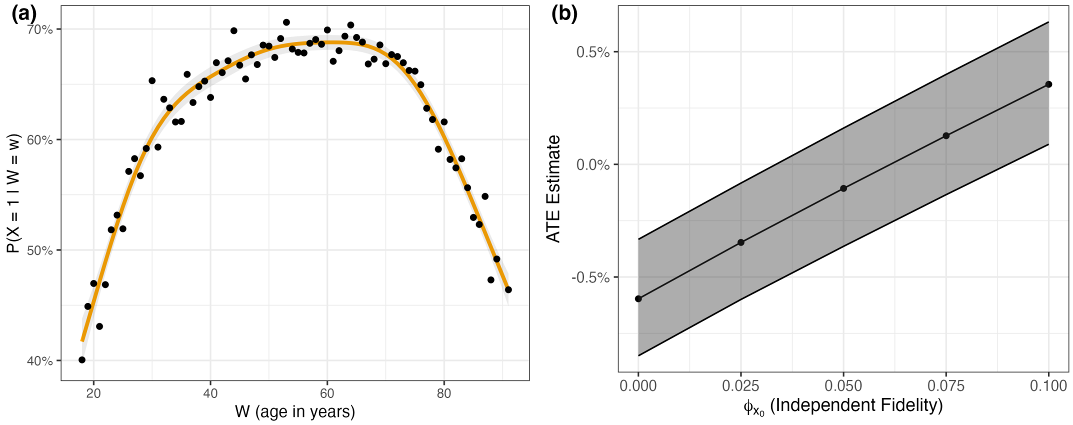


Fig 4: (a) Prevalence of overweight and obesity (X) according to age (W) $P(X = 1 \mid W = w)$, with the fit from a logistic model for X with a cubic spline term for W shown in orange; (b) x-specific independent fidelity with uncertainty quantification for increasing ϕ_{x_0}, with continuous age effects included.

on bin-counting estimators, and a parametric approach based on an exponential family model suitable for the setting of binary data. We also analyzed the statistical and computational properties of these two methods. For the non-parametric approach, the provided bound for the rate of convergence is exponential in the dimension of the confounders, reflecting the difficulty of solving the underlying inverse problem in full generality. Therefore, using the non-parametric approach is suggested for smaller-dimensional problems, allowing one to place no assumptions on the underlying distribution, while the parametric approach is suitable for larger dimensions and in presence of additional continuous confounders not subject to measurement error.

Finally, we applied both of the approaches to the original application that motivated their development – the obesity paradox – a medical phenomenon in which overweight and obese patients in the intensive care unit (ICU) show better survival than their leaner counterparts. In this context, we investigate two specific types of measurement error. The first setting is when existing comorbidities may not be recorded (only false negatives are possible), and the measurement error is independent across variables (independent fidelity). The second setting is when possibly all comorbidities are recorded as not present with some probability (zero-inflation), and are recorded correctly otherwise.

Our results demonstrated that a measurement pattern in which (i) confounders are not recorded with a probability of 5% for non-obese patients and (ii) confounders are recorded perfectly for obese patients would be required to reverse the conclusion of improved survival associated with an increased BMI. For the setting of zero-inflation, the probability of measurement error (recording all comorbidities as absent), would need to be even larger, and greater than 15% for non-obese patients, while assuming perfect recording for obese patients. Intuitively, such patterns of measurement error would imply that low BMI individuals are sicker than the data shows, while the chronic health of high BMI patients is captured correctly. Interestingly, other measurement error patterns (in which the measurement error was not modified by either X or Y, or modified by Y only) did not point to clear trends. However, of different fidelity patterns, the x-specific (BMI-specific) one is possibly the least plausible in practice, since no obvious mechanism of why BMI would modify the recording of comorbidities can be hypothesized. For instance, the death outcome would be a more likely modifier of the measurement error pattern, since the occurence of an adverse outcome may be a reason to record or follow up on comorbidities. Thus, agnostic measurement error (no modification by X, Y), or errors depending on the death outcome (y-specific patterns) would be more likely to occur in practice. Still, informative x-specific measurement error of this magnitude (5%), which is not very high, cannot be ruled out with certainty. Finally, we also remark that a study design that infers the magnitude of the true measurement error pattern is possible, by following up on subsets of patients and investigating their health records to determine if comorbidities were recorded correctly.

In conclusion, based on the evidence in this manuscript, the obesity paradox may be explained by imperfectly recorded data in the electronic health records of the hospital. However, the specific patterns of measurement error that would explain away the association of increased BMI and improved survival are not very likely in practice. Therefore, causal explanations of the obesity paradox still warrant further investigation.

SUPPLEMENTARY MATERIAL

Technical Appendices: Appendix A provides the proof of Thm. 1. Appendix B provides the proof of Thm. 2. Appendix C discusses the convergence conditions of the parametric approach. Appendix D discusses the extension of the parametric approach to correctly observed continuous data. Appendix E covers key causal inference concepts. Appendix F contains experiments on semi-synthetic data. (pdf file)

R-package binsensate: R-package binsensate provides an implementation for all the methods described in this manuscript. (GNU zipped tar file)

Code Repository: The code repository can be found on this Github link. The README file therein provides details about reproducing all the findings of the manuscript.

MIMIC-IV Data: The MIMIC-IV data used in this manuscript requires approval from the data source owners and therefore cannot be shared as a supplement to the manuscript. To access the data, please check the data source documentation. The data can be pre-processed using the script `r-utils/data-gen/data-gen.R` from the project repository.

ACKNOWLEDGEMENTS

The author would like to thank Christoph Schultheiss for valuable comments that helped improve the manuscript.

REFERENCES

Armijo, L. (1966). Minimization of functions having lipschitz continuous first partial derivatives. *Pacific Journal of Mathematics*, 16(1):1–3.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the em algorithm: From population to sample-based analysis. *The Annals of Statistics*, 45(1):77–120.

Banack, H. R. and Kaufman, J. S. (2014). The obesity paradox: understanding the effect of obesity on mortality among individuals with cardiovascular disease. *Preventive medicine*, 62:96–102.

Bareinboim, E., Correa, J. D., Ibeling, D., and Icard, T. (2022). On pearl's hierarchy and the foundations of causal inference. In *Probabilistic and Causal Inference: The Works of Judea Pearl*, page 507–556. Association for Computing Machinery, New York, NY, USA, 1st edition.

Bennett, N., Plečko, D., Ukor, I.-F., Meinshausen, N., and Bühlmann, P. (2023). ricu: R's interface to intensive care data. *GigaScience*, 12:giad041.

Berrington de Gonzalez, A., Hartge, P., Cerhan, J. R., Flint, A. J., Hannan, L., MacInnis, R. J., Moore, S. C., Tobias, G. S., Anton-Culver, H., Freeman, L. B., et al. (2010). Body-mass index and mortality among 1.46 million white adults. *New England Journal of Medicine*, 363(23):2211–2219.

Bosello, O. and Vanzo, A. (2021). Obesity paradox and aging. *Eating and Weight Disorders-Studies on Anorexia, Bulimia and Obesity*, 26(1):27–35.

Buonaccorsi, J. P. (2010). *Measurement error: models, methods, and applications*. Chapman and Hall/CRC.

Calvetti, D. and Somersalo, E. (2018). Inverse problems: From regularization to bayesian inference. *Wiley Interdisciplinary Reviews: Computational Statistics*, 10(3):e1427.

Carroll, R. J., Ruppert, D., Stefanski, L. A., and Crainiceanu, C. M. (2006). *Measurement error in nonlinear models: a modern perspective*. Chapman and Hall/CRC.

Castro, A. V. B., Kolka, C. M., Kim, S. P., and Bergman, R. N. (2014). Obesity, insulin resistance and comorbidities–mechanisms of association. *Arquivos Brasileiros de Endocrinologia & Metabologia*, 58:600–609.

Centers for Disease Control and Prevention (2020). Adult obesity facts. Accessed on 2023-08-25.

Chambers, J. M. and Hastie, T. J. (2017). Statistical models. In *Statistical models in S*, pages 13–44. Routledge.

Charlson, M. E., Pompei, P., Ales, K. L., and MacKenzie, C. R. (1987). A new method of classifying prognostic comorbidity in longitudinal studies: development and validation. *Journal of chronic diseases*, 40(5):373–383.

Cui, Y., Pu, H., Shi, X., Miao, W., and Tchetgen Tchetgen, E. (2024). Semiparametric proximal causal inference. *Journal of the American Statistical Association*, 119(546):1348–1359.

Dai, H., Ng, I., Luo, G., Spirtes, P., Stojanov, P., and Zhang, K. (2024). Gene regulatory network inference in the presence of dropouts: a causal view. *arXiv preprint arXiv:2403.15500*.

Decruyenaere, A., Steen, J., Colpaert, K., Benoit, D. D., Decruyenaere, J., and Vansteelandt, S. (2020). The obesity paradox in critically ill patients: a causal learning approach to a casual finding. *Critical Care*, 24:1–11.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society: series B (methodological)*, 39(1):1–22.

Duarte, G., Finkelstein, N., Knox, D., Mummolo, J., and Shpitser, I. (2024). An automated approach to causal inference in discrete settings. *Journal of the American Statistical Association*, 119(547):1778–1793.

Elixhauser, A., Steiner, C., Harris, D. R., and Coffey, R. M. (1998). Comorbidity measures for use with administrative data. *Medical Care*, 36(1):8–27.

Fantuzzi, G. (2005). Adipose tissue, adipokines, and inflammation. *Journal of Allergy and clinical immunology*, 115(5):911–919.

Flegal, K. M., Graubard, B. I., Williamson, D. F., and Gail, M. H. (2005). Excess deaths associated with underweight, overweight, and obesity. *Jama*, 293(15):1861–1867.

Fuller, W. A. (2009). *Measurement error models*. John Wiley & Sons.

Groetsch, C. W. and Groetsch, C. (1993). *Inverse problems in the mathematical sciences*, volume 52. Springer.

Hainer, V. and Aldhoon-Hainerová, I. (2013). Obesity paradox does exist. *Diabetes care*, 36(Suppl 2):S276.

Haslam, D., Sattar, N., and Lean, M. (2006). Obesity—time to wake up. *Bmj*, 333(7569):640–642.

Hutagalung, R., Marques, J., Kobylka, K., Zeidan, M., Kabisch, B., Brunkhorst, F., Reinhart, K., and Sakr, Y. (2011). The obesity paradox in surgical intensive care unit patients. *Intensive care medicine*, 37:1793–1799.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2020). Mimic-iv.

Johnson, A., Bulgarelli, L., Pollard, T., Horng, S., Celi, L. A., and Mark, R. (2021). Mimic-iv (version 1.0). PhysioNet.

Kuroki, M. and Pearl, J. (2014). Measurement bias and effect restoration in causal inference. *Biometrika*, 101(2):423–437.

Le Gall, J.-R., Lemeshow, S., and Saulnier, F. (1993). A new simplified acute physiology score (saps ii) based on a european/north american multicenter study. *Jama*, 270(24):2957–2963.

Louis, T. A. (1984). Estimating a population of parameter values using bayes and empirical bayes methods. *Journal of the American Statistical Association*, 79(386):393–398.

Marengoni, A., Angleman, S., Melis, R., Mangialasche, F., Karp, A., Garmen, A., Meinow, B., and Fratiglioni, L. (2011). Aging with multimorbidity: a systematic review of the literature. *Ageing research reviews*, 10(4):430–439.

Mohan, K. and Pearl, J. (2021). Graphical models for processing missing data. *Journal of the American Statistical Association*, 116(534):1023–1037.

Mullen, J. T., Moorman, D. W., and Davenport, D. L. (2009). The obesity paradox: body mass index and outcomes in patients undergoing nonbariatric general surgery. *Annals of surgery*, 250(1):166–172.

Nabi, R., Bhattacharya, R., and Shpitser, I. (2020). Full law identification in graphical models of missing data: Completeness results. In *International Conference on Machine Learning*, pages 7153–7163. PMLR.

Pearl, J. (1995). Causal diagrams for empirical research. *Biometrika*, 82(4):669–688.

Pearl, J. (2000). *Causality: Models, Reasoning, and Inference*. Cambridge University Press, New York. 2nd edition, 2009.

Pearl, J. (2018). Does obesity shorten life? or is it the soda? on non-manipulable causes. *Journal of Causal Inference*, 6(2):20182001.

Plečko, D., Bennett, N., Mårtensson, J., and Bellomo, R. (2021). The obesity paradox and hypoglycemia in critically ill patients. *Critical Care*, 25:1–15.

Rothman, K. J., Greenland, S., Lash, T. L., et al. (2008). *Modern epidemiology*, volume 3. Wolters Kluwer Health/Lippincott Williams & Wilkins Philadelphia.

Schelbert, K. B. (2009). Comorbidities of obesity. *Primary Care: Clinics in Office Practice*, 36(2):271–285.

Tilg, H. and Moschen, A. R. (2006). Adipocytokines: mediators linking adipose tissue, inflammation and immunity. *Nature reviews immunology*, 6(10):772–783.

Tremblay, A. and Bandi, V. (2003). Impact of body mass index on outcomes following critical care. *Chest*, 123(4):1202–1207.

VanderWeele, T. J. and Ding, P. (2017). Sensitivity analysis in observational research: introducing the e-value. *Annals of internal medicine*, 167(4):268–274.

Vincent, J.-L., Moreno, R., Takala, J., Willatts, S., De Mendonça, A., Bruining, H., Reinhart, C., Suter, P., and Thijs, L. G. (1996). The sofa (sepsis-related organ failure assessment) score to describe organ dysfunction/failure.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

World Health Organization (2023). Global health observatory data repository. Accessed on 2021-11-29.

Yeo, H. J., Kim, T. H., Jang, J. H., Jeon, K., Oh, D. K., Park, M. H., Lim, C.-M., Kim, K., and Cho, W. H. (2023). Obesity paradox and functional outcomes in sepsis: A multicenter prospective study. *Critical Care Medicine*, 51(6):742.