Methods of Psychological Research Online 2000, Vol.5, No.3
Internet: http://www.mpr-online.de
Institute for Science Education

Causal Regression Models II: Unconfoundedness and Causal Unbiasedness[1]

By Rolf Steyer[2],
Siegfried Gabler[3],
Alina A. von Davier[4],
and
Christof Nachtigall[4]

Abstract

We consider regression models with discrete units and a discrete treatment variable. In this framework, individual and average causal effects as well as causal unbiasedness of conditional expected values $E(Y \mid X = x)$ and of their differences were defined in a previous paper where it was also noted that a hypothesis of causal unbiasedness is not empirically testable outside the randomized experiment. Therefore, we study a stronger causality criterion which we call "unconfoundedness". To our knowledge, this is the weakest empirically testable condition implying causal unbiasedness of the conditional expected values $E(Y \mid X = x)$. Unconfoundedness holds in randomized experiments, but it may hold in nonrandomized experiments, as well. We derive theorems about sufficient and necessary conditions, about sufficient conditions, and about necessary conditions for unconfoundedness. The latter identify the hypotheses to be tested in nonrandomized experiments when it comes to testing the weakest empirically testable sufficient condition for conditional expected values $E(Y \mid X = x)$ to be causally unbiased.

***Keywords*:** Causality; Confounding; Regression Models; Randomization; Rubin´s Approach to Causality

---

[1] We would like to thank Albrecht Iseler (Free University Berlin) for a discussion in which the idea for some parts of this paper has been developed. We would also like to extend our gratitude to Angelika von der Linde (University of Edinburgh), to Donald B. Rubin (Harvard University), and to Peter Kischka (FSU Jena) who all gave helpful critical comments on previous versions of this paper.
[2] Address for correspondence: Prof. Dr. Rolf Steyer, Friedrich-Schiller-Universität Jena, Am Steiger 3 – Haus 1, D-07743 Jena, Germany. Email: rolf.steyer@uni-jena.de
[3] Center for Surveys, Methods, and Analyses (ZUMA), Mannheim, Germany
[4] Friedrich-Schiller-University Jena, Germany

In a previous paper (Steyer, Gabler, von Davier, Nachtigall, & Buhl, 2000) we reformulated the theory of individual and average causal effects in terms of classical probability theory and illustrated it by some examples. This theory goes back to Neyman (1923/1990; 1935) and has been adopted and enriched by Rubin (1973a, b, 1974, 1977, 1978, 1985, 1986, 1990), Holland and Rubin (1983), Holland (1986, 1988a, b), Rosenbaum and Rubin (1983a, b, 1984, 1985a, b), Rosenbaum (1984a, b, c), and Sobel (1994, 1995), for instance. More specifically, we introduced a probabilistic framework: *a potential causal regression model with discrete units and discrete treatment variable*[5], which consists of three components:

- a probability space $(\Omega, \mathfrak{A}, P)$, which represents the random experiment, i.e., the empirical phenomenon considered,
- a regression $E(Y|X)$, the potential causal interpretation of which is focused, where X represents a discrete (but not necessarily univariate) treatment variable and Y a real-valued (and not necessarily continuous) response variable,
- a nonnumeric random variable U, the value of which is the observational unit drawn.

The probability space is chosen such that X, Y, and U are random variables on $(\Omega, \mathfrak{A}, P)$, i.e., X, Y, and U have a joint distribution. In this framework and notation, we presented the basic concepts of the theory of individual and average causal effects: the *individual conditional expected values* $E(Y|X=x, U=u)$, the *individual causal effects* $ICE_u(i, j) := E(Y|X=x_i, U=u) - E(Y|X=x_j, U=u)$ of a treatment x_i vs. a treatment x_j, the *causally unbiased expected value* [denoted $CUE(Y|X=x)$], and the *average causal effect* $ACE(i, j)$ (see Figure 1 of Steyer et al., 2000). We also showed how these concepts are related to the conditional expected values $E(Y|X=x)$ and the *prima facie effects* $PFE(i, j) = E(Y|X=x_i) - E(Y|X=x_j)$ that are usually focused in statistical estimation and hypothesis testing. Specifically, it has been shown that both, (a) stochastic independence of U and X (that can be created via random assignment of units to treatment conditions) as well as (b) unit-treatment homogeneity, [i.e., $E(Y|X, U) = E(Y|X)$] imply the equations $E(Y|X=x) = CUE(Y|X=x)$], and $PFE(i, j) = ACE(i, j)$ [i.e., *causal unbiasedness* of $E(Y|X=x)$ and of $PFE(i, j)$].

We argued that the theory of individual and average causal effects reviewed above has some major limitations. The *first* one is that a proposition about causal unbiasedness is not empirically falsifiable: Postulating that the prima facie effect $PFE(i, j)$ is equal to the average causal effect $ACE(i, j)$ or that the conditional expected values $E(Y|X=x)$ are equal to the causally unbiased expected values $CUE(Y|X=x)$ of Y given x does not imply anything one could show to be wrong in an empirical appli-

[5] For a more general formal framework, see Steyer (1992).

cation. The reason is that the computation of the $ACE(i, j)$ and the $CUE(Y \mid X = x)$ involve *all* individual conditional expected values $E(Y \mid X = x, U = u)$ in the population and that it is not possible to estimate all of them. (See the fundamental problem of causal inference in Holland, 1986 or Steyer et al., 2000).

The *second limitation* is that even if both $PFE(i, j) = ACE(i, j)$ and $E(Y \mid X = x) = CUE(Y \mid X = x)$ hold in the total population, the corresponding equations may not hold in the subpopulations, for instance, in the subpopulations of males and females. Hence, causal unbiasedness of the prima facie effects in the subpopulations might not hold although being correct in the total population.

Because of the nonfalsifiability we argued that, although the theory of individual and average causal effects provides indispensable concepts, it is not really complete as a methodological basis for causal modeling *outside* the randomized experiment in which treatment assignment may not be random. Completing this methodological basis is the central goal of this paper. Specifically, we aim at enriching the theory of individual and average causal effects by introducing the concept of *unconfoundedness* of a regression $E(Y \mid X)$. Unconfoundedness will be defined such that:

(a) a proposition that a regression $E(Y \mid X)$ is unconfounded is empirically falsifiable,

(b) unconfoundedness of $E(Y \mid X)$ implies that the regressions $E_{W=w}(Y \mid X)$ in the subpopulations, are causally unbiased as well.[6]

Among all criteria fulfilling these requirements we prefer that criterion which is logically the weakest, i.e., which is implied by the others, but does not imply one of the others.

If requirement (a) were not fulfilled, everybody could make causal propositions, nobody would have a chance to falsify it. Requirement (b) means that causal interpretations such as causal unbiasedness should be transferable from the total population into its subpopulations and in this sense be "stable" or "invariant".

Hence, the goal of this paper is to introduce a causality criterion that fulfills the requirements (a) and (b) mentioned above, compare it to some other criteria, illustrate it by an example and study its sufficient and necessary conditions. More specifically, we will present and discuss different causality criteria in section 1, including *strong ignorability* and *unconfoundedness*. The latter is, according to requirements (a) and (b), the most favorable one. In section 2 we will illustrate the different criteria by an example. Section 3 is devoted to a necessary and sufficient condition of unconfoundedness. Next we will study sufficient conditions of unconfoundedness (section 4), and then sufficient conditions of strong ignorability (section 5). Then we turn to the necessary conditions of

[6] Each value w of W represents a subpopulation.

unconfoundedness (section 6). One of these necessary conditions is causal unbiasedness of $E(Y|X)$. Section 7 focusses two general procedures to falsify unconfoundedness, which will be illustrated by an example in section 8. Section 9 deals with necessary conditions for unconfoundedness under special assumptions such as linearity and W-conditional linearity of $E(Y|X, W)$, where W is a variable each value of which represents a subpopulation. Finally, the merits and limits of the theory presented are discussed in section 10.

1. Some Causality Criteria

The weakest criterion we consider is causal unbiasedness of *the treatment regression* $E(Y|X)$ in the total population, i.e.,

$$E(Y|X=x) = CUE(Y|X=x), \qquad \text{for each value } x \text{ of } X, \tag{1}$$

with

$$CUE(Y|X=x) := \sum_u (Y|X=x, U=u)\ P(U=u), \tag{2}$$

where the summation is across all observational units u in the (total) population $\mathbf{\Omega}_U$ (i.e., the set of all units considered). This term has been defined to be the *causally unbiased expected value* of Y given x (Steyer et al., 2000). Equation (1) means that each $E(Y|X=x)$ is equal to the expected value of the individual expected values $E(Y|X=x, U=u)$ across the distribution of the observational units. Equation (1) is sufficient to imply $PFE(i, j) = ACE(i, j)$, i.e., causal unbiasedness of the prima facie effects in the total population (see Steyer et al., 2000). The equation $E(Y|X=x) = CUE(Y|X=x)$, defines *causal unbiasedness of* a conditional expected value $E(Y|X=x)$.

From the perspective of the theory of individual and average causal effects, there is no doubt that Equation (1), and with it, the concept of causal unbiasedness, is a desirable and indispensable property of the conditional expected values $E(Y|X=x)$ in meaningful substantive applications. However, as mentioned before, Equation (1) neither implies anything that could be falsified in an empirical application [see requirement (a)] nor does it imply unbiasedness of the regressions $E_{W=w}(Y|X)$ in the subpopulations represented by $W = w$ [see requirement (b)]. Hence, Equation (1) [*causal unbiasedness of the regression* $E(Y|X)$] is too weak to qualify as a satisfactory causality criterion applicable to experimental studies with nonrandom assignment.[7] The same ar-

[7] In a similar vein Pearl (1998) argues that unbiasedness might be incidental which necessitates the need for "stable unbiasedness" (see also Example III in Steyer et al., 2000).

gument applies, of course, to the criterion $PFE(i, j) = ACE(i, j)$, *causal unbiasedness of the prima facie effects.* Both criteria lack empirical falsifiability.

As a second criterion we consider "strong ignorability" described, e.g., by Rosenbaum and Rubin (1983a). They defined potential response variables Y_i: $\Omega \to$ IR for each treatment condition x_i. In our notation we could define these variables for each value x_i by $Y_i(\omega) = f(u) = E(Y \mid X = x_i, U = u)$, for all values u of U. Defined in this way, it is obvious that the variables Y_i, $i = 1, \ldots, n_x$, are functions of U and therefore have a joint distribution with X, because U and X have joint distributions.[8] Hence, stochastic independence of U and X implies that the variables $Y_1, \ldots, Y_{n_x}$ and X are stochastically independent (denoted $Y_1, \ldots, Y_{n_x} \perp X$). The condition (a) $Y_1, \ldots, Y_{n_x} \perp X$ and (b) $0 < P(X = x_i) < 1$, for each value x_i of X has been called *strong ignorability* (see, e.g., Rosenbaum & Rubin, 1983a, p. 213).[9] Strong ignorability implies causal unbiasedness of the conditional expected values and the prima facie effects. However, it is neither empirically testable, and it is unknown whether or not it implies causal unbiasedness in the subpopulations. (See Note 1 in Appendix A).

Let us now discuss several other sufficient conditions for causal unbiasedness that *are* empirically falsifiable. A *third* criterion we might consider is *stochastic independence* of X and U. This criterion, in fact, fulfills both requirements discussed above. It *is* empirically falsifiable and it implies causal unbiasedness of the conditional expected values $E(Y \mid X = x, W = w)$ in the subpopulations represented by $W = w$ (see Th. 3 and Th. 6). However, Theorem 2 of Steyer et al. (2000) shows that there is a *fourth* criterion that is also falsifiable and implies causal unbiasedness of the conditional expected values $E(Y \mid X = x, W = w)$: *unit-treatment homogeneity*, i.e., $E(Y \mid X, U) = E(Y \mid X)$. Hence, independence of U and X would be unnecessarily strong.

The *fifth* criterion to be considered is "X and U are independent *or* $E(Y \mid X, U) = E(Y \mid X)$".[10] This would also fulfill both requirements discussed above. In fact, this criterion is already very close to our favorite one which is somewhat weaker and still fulfills the two requirements (a) and (b).

The *sixth criterion* will be called *unconfoundedness* of the regression $E(Y \mid X)$.[11] It is defined as follows. Note that throughout the paper we will presume that $\langle (\Omega, \mathfrak{A}, P), E(Y \mid X), U \rangle$ is a potential causal regression model.

---

[8] In Rubin's notation, the causally unbiased expected values would be denoted $E(Y_i)$, i.e., $E(Y_i) = CUE(Y \mid X = x_i)$.

[9] For simplicity, we restrict our discussion to the case where there is no concomitant variable or covariate. Rosenbaum and Rubin consider the case with a covariate. Also note that strong ignorability implies $E(Y_i) = E(Y_i \mid X = x)$ for each pair (i, x) of indices $i \in \{1, \ldots, n_x\}$ and values x of X.

[10] This criterion is very close to what has been proposed by Pearl (1998) in his definitions 2 and 3.

[11] In previous papers (see e.g. Steyer, Gabler, & Rucai, 1996 and Steyer, von Davier, Gabler, & Schuster, 1997), we have defined unconfoundedness in a different but equivalent way (see Th. 2).

Definition 1. *$E(Y \mid X)$ is called unconfounded if for each value x of X:*

$$P(X = x \mid U = u) = P(X = x) \qquad \textit{for each value } u \textit{ of } U \tag{3}$$

or

$$E(Y \mid X = x,\ U = u) = E(Y \mid X = x) \qquad \textit{for each value } u \textit{ of } U. \tag{4}$$

This sixth criterion differs from the fifth one by the fact that "Equation (3) or Equation (4)" is postulated *within each value x of X*. (For an example, see Table 2.) Hence, *within each treatment condition* we require equal treatment assignment probabilities for each unit u [see Eq. (3)] or homogeneity of the units. Only *one* of these equations has to be true within each treatment condition x. In contrast, the fifth criterion requires Equation (3) to be true for *all* treatment conditions x or Equation (4) to be true for *all* treatment conditions x. Whereas from a substantive point of view this difference seems negligible, it is important from a mathematical point of view: the two criteria are *not* equivalent to each other. In fact, the fifth criterion implies the sixth criterion, but not vice versa. Table 1 displays a summary of the causality criteria discussed above.

Table 1. Summary of the causality criteria discussed in the paper

1. Causal unbiasedness of $E(Y \mid X)$	$E(Y \mid X = x) := \sum_u E(Y \mid X = x,\ U = u)\ P(U = u)$, for each value x of X
2. Strong ignorability	$Y_1, \ldots, Y_{n_x} \perp X \quad$ *and* $\quad 0 < P(X = x_i) < 1$
3. Independence of X and U	$X \perp U$
4. Unit-treatment homogenity	$E(Y \mid X,\ U) = E(Y \mid X)$
5. Independence of X and U *or* unit-treatment homogeneity	$X \perp U \quad$ *or* $\quad E(Y \mid X,\ U) = E(Y \mid X)$
6. Unconfoundedness of $E(Y \mid X)$	For each value x of X: $P(X = x \mid U = u) = P(X = x)$ for each value u of U or $E(Y \mid X = x,\ U = u) = E(Y \mid X = x)$ for each value u *of* U

2. An Example

Table 2 gives an example illustrating the criteria discussed above. The example is constructed in such a way that unconfoundedness and causal unbiasedness but none of the other causality criteria hold. We consider three treatment conditions x_1 to x_3 and a

population of eight units. We assume that each unit has the probability $P(U = u) = 1/8$ to be sampled. Let us first look at treatment x_1. Column 4 displays the individual assignment probabilities for treatment x_1. They are all *the same*, namely 1/2, for each individual unit. For this specific treatment condition x_1 we have *different* individual conditional expected values $E(Y \mid X = x_1,\ U = u)$ (see column 5). For treatments x_2 and x_3 things are different. Column 6 contains *different* treatment assignment probabilities $P(X = x_2 \mid U = u)$ for treatment condition x_2, and column 7 displays equal individual conditional expected values $E(Y \mid X = x_2,\ U = u)$. The same is true for treatment condition x_3. Again, we have different treatment assignment probabilities $P(X = x_3 \mid U = u)$ but equal individual conditional expected values $E(Y \mid X = x_3,\ U = u)$.

In order to check causal unbiasedness, we may first compute the three conditional expected values $E(Y \mid X = x_1) = 115$, $E(Y \mid X = x_2) = 105$, and $E(Y \mid X = x_3) = 110$ from Equation

$$E(Y \mid X = x) := \sum_u E(Y \mid X = x,\ U = u)\ P(U = u \mid X = x), \tag{5}$$

which always holds for a discrete random variable U. Using

$$P(U = u \mid X = x) = P(X = x \mid U = u)\ P(U = u)\ /\ P(X = x) \tag{6}$$

yields $P(U = u \mid X = x_1) = (1/2) \cdot (1/8)\ /\ (1/2) = 1/8$ for each unit u and

$$E(Y \mid X = x_1) = 82 \cdot 1/8 + \ldots + 152 \cdot 1/8 = 115.$$

For the second treatment condition x_2 the corresponding formulas lead to $P(U = u \mid X = x_2) = (1/10) \cdot (1/8)\ /\ (1/4) = 1/20$ for u_1 and u_2, to $P(U = u \mid X = x_2) = (2/10) \cdot (1/8)\ /\ (1/4) = 2/20$ for u_3 and u_4, to $P(U = u \mid X = x_2) = (3/10) \cdot (1/8)\ /\ (1/4) = 3/20$ for u_5 and u_6, and to $P(U = u \mid X = x_2) = (4/10) \cdot (1/8)\ /\ (1/4) = 4/20$ for u_7 and u_8. Hence, Equation (5) now yields

$$E(Y \mid X = x_2) = 105 \cdot (1/20 + 1/20 + 2/20 + 2/20 + 3/20 + 3/20 + 4/20 + 4/20) = 105.$$

Analogously, we can compute the conditional expected value $E(Y \mid X = x_3) = 110$.

Table 2. An example in which the treatment regression $E(Y \mid X)$ is unconfounded and causally unbiased but none of the other causality criteria discussed in this paper hold

Observational-unit variable U	$P(U = u)$	W (gender)	$P(X = x_1 \mid U = u)$	$E(Y \mid X = x_1, U = u)$	$P(X = x_2 \mid U = u)$	$E(Y \mid X = x_2, U = u)$	$P(X = x_3 \mid U = u)$	$E(Y \mid X = x_3, U = u)$
u_1	1/8	m	1/2	82	1/10	105	4/10	110
u_2	1/8	m	1/2	89	1/10	105	4/10	110
u_3	1/8	m	1/2	101	2/10	105	3/10	110
u_4	1/8	m	1/2	108	2/10	105	3/10	110
u_5	1/8	f	1/2	118	3/10	105	2/10	110
u_6	1/8	f	1/2	131	3/10	105	2/10	110
u_7	1/8	f	1/2	139	4/10	105	1/10	110
u_8	1/8	f	1/2	152	4/10	105	1/10	110

Note: The (unconditional) probabilities for the three treatments are $P(X = x_1) = 1/2$, $P(X = x_2) = P(X = x_3) = 1/4$.

In order to check causal unbiasedness, the *conditional expected values* computed above have to be compared to the *causally unbiased conditional expected values*, which may be computed by Equation (2). For the treatment condition x_1 this equation yields

$$CUE(Y \mid X = x_1) := \sum_u E(Y \mid X = x_1, U = u)\, P(U = u)$$

$$= (82 + 89 + 101 + 108 + 118 + 131 + 139 + 152) \cdot 1/8 = 115,$$

for treatment condition x_2

$$CUE(Y \mid X = x_2) := \sum_u E(Y \mid X = x_2, U = u)\, P(U = u) = (105 + \ldots + 105) \cdot 1/8 = 105,$$

and for treatment condition x_3

$$CUE(Y \mid X = x_3) := \sum_u E(Y \mid X = x_3, U = u)\, P(U = u) = (110 + \ldots + 110) \cdot 1/8 = 110.$$

Hence, in this example, the conditional expected values are unbiased, i.e., $E(Y \mid X = x) = CUE(Y \mid X = x)$ for each value x of X, and the prima facie effects $PFE(i, j) = E(Y \mid X = x_i) - E(Y \mid X = x_j)$ are unbiased as well, i.e., $PFE(i, j) = ACE(i, j)$, for each pair of treatment conditions x_i and x_j.

We now check the other causality criteria discussed in the last section. The second causality criterion, *strong ignorability*, does *not* hold in this example. The independence condition $Y_1, \ldots, Y_{n_x} \perp X$ implies the conditional expected values $E(Y_1 \mid X = x_1)$ and $E(Y_1 \mid X = x_2)$ to be identical. However, in our example:

$$E(Y_1 \mid X = x_1) = \sum_u Y_1(u)\, P[Y_1 = Y_1(u) \mid X = x_1] = \sum_u Y_1(u)\, P(U = u \mid X = x_1) =$$

$$= \sum_u \left[Y_1(u)\, [P(X = x_1 \mid U = u) \cdot P(U = u) \,/\, P(X = x_1)] \right]$$

$$= 82 \cdot 1/8 + \ldots + 152 \cdot 1/8 = 115$$

and

$$E(Y_1 \mid X = x_2) = \sum_u Y_1(u)\, P[Y_1 = Y_1(u) \mid X = x_2] = \sum_u Y_1(u)\, P(U = u \mid X = x_2) =$$

$$= \sum_u \left[Y_1(u)\, [P(X = x_2 \mid U = u) \cdot P(U = u) \,/\, P(X = x_2)] \right]$$

$$= (82 + 89) \cdot 1/20 + (101 + 108) \cdot 2/20 + (118 + 131) \cdot 3/20 + (139 + 152) \cdot 4/20 = 125.$$

Hence, we can conclude that strong ignorability does not hold in this example.

The third criterion, *stochastic independence* of X and U does not hold, because, unlike for treatment condition x_1, the individual assignment probabilities $P(X = x_2 \mid U = u)$ and $P(X = x_3 \mid U = u)$ are not the same for each observational unit u for treatment conditions x_2 and x_3.

The fourth causality criterion, *unit-treatment homogeneity* also does not hold. Although there are equal individual conditional expected values $E(Y \mid X = x, U = u)$ in treatment conditions x_2 and x_3, these values differ in treatment condition x_1.

The fifth causality criterion, *stochastic independence of X and U or unit-treatment homogeneity* does not hold as well. However, the sixth causality criterion, unconfoundedness *does* hold. In each treatment condition x we have either equal assignment probabilities $P(X = x \mid U = u)$ (for x_1) or equal individual conditional expected values $E(Y \mid X = x, U = u)$ (for x_2 and x_3).

3. Some Equivalent Formulations of Unconfoundedness

We will now study several conditions which are equivalent to unconfoundedness as defined above. These conditions will provide the theoretical basis for empirical falsifiability. All these conditions involve random variables W for which there exists a map-

ping f such that $W = f(U)$ is the composition of U with f. As can easily be seen, such a variable W partitions the population $\mathbf{\Omega}_U$ into subpopulations (such as the set of *male* and the set of *female* persons). Presuming a potential causal regression model $\langle(\mathfrak{Q}, \mathfrak{A}, P),\ E(Y|X),\ U\rangle$ and using these variables, we may formulate the following theorem.[12]

Theorem 1. *$E(Y \mid X)$ is unconfounded if and only if for each value x of X and for each $W = f(U)$*

$$P(X = x|\ W = w) = P(X = x) \qquad \textit{for each value } w \textit{ of } W \tag{7}$$

or

$$E(Y|\ X = x,\ W = w) = E(Y|\ X = x) \qquad \textit{for each value } w \textit{ of } W. \tag{8}$$

According to this theorem, unconfoundedness postulates, *within* each treatment condition x, equal treatment probabilities for each subpopulation w or equal conditional expected values across all subpopulations (represented by the different values w of W). Note that, in empirical applications, this theorem already provides a basis for falsification of unconfoundedness, because, given a specific $W = f(U)$, each term in these two equations is empirically estimable. At least one of the two equations above has to be true for each $W = f(U)$ and a given value x of X. If neither Equation (7) nor Equation (8) holds for a given $W = f(U)$, we can conclude that $E(Y \mid X)$ is *confounded*.

The next theorem will show the surprising fact that unconfoundedness as defined by the sixth criterion (see Def. 1) is equivalent to unconfoundedness as defined by Steyer, Gabler, and Rucai (1996).[13]

Theorem 2. *$E(Y|X)$ is unconfounded if and only if for each $W = f(U)$*

$$E(Y|\ X = x) = \sum_w E(Y|\ X = x,\ W = w)\ P(W = w), \qquad \textit{for each value } x \textit{ of } X. \tag{9}$$

Note that the summation is across all values w of W. According to this theorem, postulating Equation (9) for *each* variable $W = f(U)$ is equivalent to *(general) unconfoundedness of the regression* $E(Y|X)$. Also note that Equation (9) is *not* equivalent with Equation (7) or Equation (8). Instead Theorem 2 formulates (via Th. 1) the equivalence between "Equation (9) for each $W = f(U)$" and "Equation (7) or Equation (8) for each $W = f(U)$". According to Equation (9) unconfoundedness implies that each conditional expected value $E(Y|X = x)$ is the average of the corresponding ex-

[12] For proofs see Appendix B.

[13] This formulation of unconfoundedness has also been called *weak causality* by Steyer (1992). There one will also find a formulation which is not restricted to discrete treatments and a discrete unit variable.

pected values $E(Y \mid X = x,\ W = w)$ in the subpopulations represented by the values w of W.

Corollary 1. *If the treatment regression $E(Y \mid X)$ is unconfounded, then for each $W = f(U)$:*

$$PFE(i, j) = \sum_w [E(Y \mid X = x_i,\ W = w) - E(Y \mid X = x_j,\ W = w)]\ P(W = w). \qquad (10)$$

According to this corollary, unconfoundedness implies that each prima facie effect, i.e., each difference $E(Y \mid X = x_i) - E(Y \mid X = x_j)$ between two values of the regression $E(Y \mid X)$, is the average of the differences $E(Y \mid X = x_i,\ W = w) - E(Y \mid X = x_j,\ W = w)$ *across all subpopulations* represented by different values w of W.

Also note that Theorem 2 holds irrespective of any specific parameterization of the regression $E(Y \mid X)$, be it linear or not. For instance, if Y is dichotomous with values 0 and 1 and X is numerical, Theorem 2 also holds for a logistic regression $E(Y \mid X) = P(Y = 1 \mid X) = \exp(\gamma_0 + \gamma_1 X) / [1 + \exp(\gamma_0 + \gamma_1 X)]$.

One might expect that Equation (9) is always true if W is discrete. However, Example Ia of Steyer et al. (2000) (with $W = U$) shows that this expectation is wrong. An equation for the conditional expected values $E(Y \mid X = x)$ that *is* always true if W is discrete is:

$$E(Y \mid X = x) = \sum_w E(Y \mid X = x,\ W = w)\ P(W = w \mid X = x), \text{ for each value } x \text{ of } X. \qquad (11)$$

Note that Equation (9) also provides a basis for empirical falsification of unconfoundedness, because each term in this equation is empirically estimable. Hence, both Theorems 1 and 2 provide possibilities for falsifying unconfoundedness.

It will be useful to supplement the general concept of unconfoundedness by the concept of unconfoundedness *with respect to a specific variable* W. The definition will be such that general unconfoundedness is equivalent to unconfoundedness with respect to each $W = f(U)$.

Definition 2. *$E(Y \mid X)$ is called unconfounded with respect to $W = f(U)$ if Equation (9) holds for each value x of X.*

What has been achieved so far? We have three different but equivalent ways of formulating unconfoundedness (see Table 3). This gives us a deeper understanding of this concept. Specifically, we have shown that, in contrast to causal unbiasedness and strong ignorability, unconfoundedness is empirically falsifiable (see versions 2 and 3 in Table 3).

Table 3. Three equivalent formulations of unconfoundedness

1.	For each value x of X: $P(X = x \mid U = u) = P(X = x)$ for each value u of U or $E(Y \mid X = x,\ U = u) = E(Y \mid X = x)$ for each value u *of* U
2.	For each variable $W = f(U)$ and for each value x of X: $P(X = x \mid W = w) = P(X = x)$ for each value w of W or $E(Y \mid X = x,\ W = w) = E(Y \mid X = x)$ for each value w *of* W
3.	For each variable $W = f(U)$ and for each value x of X: $E(Y \mid X = x) := \sum_w E(Y \mid X = x,\ W = w)\ P(W = w)$, for each value x of X

4. Sufficient Conditions for Unconfoundedness

Contrasting unconfoundedness to other criteria, we will now show more formally that unconfoundedness is a comparatively weak sufficient condition for causal unbiasedness. This will be achieved studying some sufficient conditions for a treatment regression $E(Y \mid X)$ to be unconfounded. The sufficient conditions will also help to understand the concept and how it is related to the experiment with random assignment of units to experimental conditions.

Theorem 3. *Each of the following conditions is sufficient for unconfoundedness of the regression* $E(Y \mid X)$*:*

(i) *U and X are stochastically independent;*

(ii) *(Unit-treatment homogeneity)* $E(Y \mid X,\ U) = E(Y \mid X)$*;*

(iii) *(Strong causality) For each* $W = f(U)$ *there exists a function h such that*

$$E(Y \mid X,\ W) = E(Y \mid X) + h(W)\ . \tag{12}$$

Note that each of these conditions is sufficient for unconfoundedness of $E(Y \mid X)$ but not necessary, i.e., each of them implies unconfoundedness of $E(Y \mid X)$, but unconfoundedness of $E(Y \mid X)$ neither implies (i), nor (ii), nor (iii). Condition (i), independence of X and U, is important for understanding the role of randomization, i.e., random assignment of units to treatment conditions. Randomization is the only way for the experimenter to deliberately create one of the sufficient conditions of unconfoundedness.

Equation (12) postulates the additive decomposability of the regression $E(Y|X,\ W)$ into a function $g\,(X)$ and a function $h\,(W)$, i.e.,

$$E(Y|X,\ W) = g\,(X)\ +\ h\,(W) \tag{13}$$

with the additional requirement $g\,(X) = E(Y|X)$, which, if Equation (13) holds, is equivalent to $E[\,h\,(W)|X\,] = E[\,h\,(W\,)]$.

For $W = U$, Equation (13) may be called *unit-treatment additivity*. It postulates the additive decomposability of the regression $E(Y|X,\ U)$ into a function $g\,(X)$ and a function $h\,(U)$, with the additional requirement $g\,(X)\ = E(Y|X)$. Equation (13) does not allow for interactions between X and W in the sense of analysis of variance (or moderator effects in terms of moderator regression models.) Strong causality as defined by condition (iii) requires invariance of the individual causal effects across all observational units, because, for $W = U$, Equation (13) implies

$$E(Y|X = x_1,\ U\ =\ u) - E(Y|X = x_2,\ U\ =\ u)$$

$$= g\,(x_1)\ +\ h\,(u) - [g\,(x_2)\ +\ h\,(u)] = g\,(x_1) - g\,(x_2)\,. \tag{14}$$

It should be noted that there is no sufficient condition for such an invariance of individual causal effects that could deliberately be created by the experimenter. Specifically, Equation (13) is not necessarily (and in fact seldom) true in the randomized experiment, in contrast, to unconfoundedness that always holds if there random assignment of units to treatment conditions. Obviously, unit-treatment homogeneity is a special case of unit-treatment additivity with $h\,(W)\ = 0$.

5. Sufficient Conditions for Strong Ignorability

In this section we state some sufficient conditions for strong ignorability. According to our next theorem, the first two sufficient conditions of unconfoundedness mentioned in Theorem 3 are also sufficient for strong ignorability.

Theorem 4. *Each of the following conditions is sufficient for strong ignorability:*

(i) *U and X are stochastically independent;*

(ii) *(Unit-treatment homogeneity)* $E(Y|X,\ U\ ) = E(Y|X)$*;*

Note that each of these conditions is *sufficient* for strong ignorability but not necessary, i.e., each of them implies strong ignorability but strong ignorability neither implies (i) nor (ii).

6. Necessary Conditions of Unconfoundedness

The first necessary condition formulated in this section relates unconfoundedness to causal unbiasedness (Th. 5). Next we show that unconfoundedness and, therefore, causal unbiasedness, can be transferred to each subpopulation. Other necessary conditions treated are important for causal modeling in *nonrandomized* experiments, because they are the logical basis for falsifying a hypothesis of unconfoundedness.

Theorem 5. *If the treatment regression $E(Y \mid X)$ is unconfounded, then each of the following propositions hold:*

(i) *$E(Y \mid X)$ is causally unbiased, i.e., $E(Y \mid X = x) = CUE(Y \mid X = x)$ for each value x of X;*

(ii) *the prima facie effects are causally unbiased, i.e., $PFE(i, j) = ACE(i, j)$ for all pairs (x_i, x_j) of values of X.*

We now turn to a very powerful theorem according to which unconfoundedness of a regression $E(Y \mid X)$ allows to conclude that unconfoundedness and, therefore, causal unbiasedness (see Th. 5), can be transferred to each subpopulation. This is not only important for the interpretation of expected values and their differences in subpopulations but also allows to conclude that all properties of an unconfounded regression also hold for the corresponding regressions within subpopulations.

Theorem 6. *If $E(Y \mid X)$ is unconfounded and $W = f(U)$, then, for each value w of W, the regression $E_{W=w}(Y \mid X)$ of Y on X in the subpopulation represented by $W = w$ is unconfounded as well.*

As mentioned before, this theorem implies that all properties of an unconfounded regression also apply for the regressions $E_{W=w}(Y \mid X)$ within the subpopulations. Instead of rewriting all propositions for the subpopulations we just consider the property of causal unbiasedness of the regressions $E_{W=w}(Y \mid X)$ of Y on X in the subpopulation represented by $W = w$.

Corollary 2. *If $E(Y \mid X)$ is unconfounded and $W = f(U)$, then the regressions $E_{W=w}(Y \mid X)$ of Y on X in the subpopulations represented by $W = w$ are causall unbiased, i.e.,*

$$E(Y \mid X = x,\ W = w) = \sum_u E(Y \mid X = x,\ U = u,\ W = w)\ P(U = u \mid W = w), \quad (15)$$

for each pair (x, w) of values of X and W.[14]

[14] Note that $E(Y \mid X = x,\ U = u) = E(Y \mid X = x,\ U = u,\ W = w)$, because we presuppose that W is a measurable function of U (see the definitions of U in Steyer et al., 2000 and of W above).

Theorem 6 and Corollary 2 show that *unconfoundedness* is a stable property in the sense that it carries over from the total population to all subpopulations. In contrast, *causal unbiasedness* may be incidentally true in the total population but fail in any subpopulation (see Example III of Steyer et al., 2000).

Table 4. Necessary conditions of unconfoundedness

1.	The regressions $E_{W=w}(Y \mid X)$ in each subpopulation $W = w$ are unconfounded and, therefore, causally unbiased
2.	The prima facie effects $PFE(i, j)$ in the total population as well as the prima facie effects $PFE_{W=w}(i, j)$ in each subpopulation are causally unbiased
3.	For each variable $W = f(U)$ and for each value x of X: $P(X = x \mid W = w) = P(X = x)$ for each value w of W or $E(Y \mid X = x,\ W = w) = E(Y \mid X = x)$ for each value w *of* W
4.	For each variable $W = f(U)$ and for each value x of X: $E(Y \mid X = x) := \Sigma_w\ E(Y \mid X = x,\ W = w)\ P(W = w)$, for each value x of X

Table 4 gives a summary of the most important necessary conditions of unconfoundedness. The first two necessary conditions are important for substantive interpretations, because they add meaning to the concept of unconfoundedness which is otherwise not easily seen. The second two necessary conditions are also sufficient. Conditions 1, 3, and 4 may be used when it comes to falsification trials of the hypothesis of unconfoundedness. It should be noticed that, *for a given variable* W "Equation (7) or (8)" is stronger than "Equation (9)", i.e., "Equation (7) or (8)" implies "Equation (9)" but not vice versa.

7. General Procedures to Falsify Unconfoundedness

An important goal of this paper is to present a theory that may serve as a methodological foundation of causal modeling also outside the randomized experiment. Since a hypothesis that the conditional expected values $E(Y \mid X = x)$ are causally unbiased is not falsifiable, we introduced unconfoundedness of $E(Y \mid X)$ as the weakest empirically falsifiable sufficient condition for causal unbiasedness. How to proceed if we want to try to falsify the hypothesis of unconfoundedness? A first general principle is to choose a variable $W = f(U)$ and compare the estimates on the right-hand side of Equation (9)

to the estimates on the right-hand side of Equation (11), i.e., to compare, for each value x of X,

$$E_{adj\ for\ W}(Y \mid X = x) := \Sigma_w E(Y \mid X = x,\ W = w)\ P(W = w) \tag{16}$$

to

$$E(Y \mid X = x) = \Sigma_w E(Y \mid X = x,\ W = w)\ P(W = w \mid X = x). \tag{17}$$

Note that the last equation is always true if W is discrete. If, in a specific empirical application, the estimate of $E_{adj\ for\ W}(Y \mid X = x)$ cannot be assumed to be an estimate of $E(Y \mid X = x)$, then unconfoundedness, too, cannot be assumed to hold.

According to the considerations above, falsifying in a specific application the hypothesis of unconfoundedness of $E(Y \mid X)$ can be achieved by falsifying the null hypothesis of unconfoundedness of $E(Y \mid X)$ *with respect to a specific* $W = f(U)$, i.e., by falsifying the hypothesis:

$$E_{adj\ for\ W}(Y \mid X = x) - E(Y \mid X = x) = 0, \qquad \text{for each value } x \text{ of } X. \tag{18}$$

This hypothesis is equivalent to Equation (9). It may also equivalently be written:

$$\Sigma_w E(Y \mid X = x,\ W = w)\ [P(W = w) - P(W = w \mid X = x)] = 0, \ \text{ for each } x \text{ of } X. \tag{19}$$

If the unconditional probabilities $P(W = w)$ and the conditional probabilities $P(W = w \mid X = x)$ are known, Equation (19) is a linear hypothesis about the parameters $E(Y \mid X = x,\ W = w)$.

A second general procedure can be based on condition 4 in Table 4. Again, choose a variable $W = f(U)$ and check if for each given value x of X condition

$$P(X = x \mid W = w) = P(X = x) \qquad \text{for each value } w \text{ of } W \tag{20}$$

holds or

$$E(Y \mid X = x,\ W = w) = E(Y \mid X = x) \qquad \text{for each value } w \textit{ of } W. \tag{21}$$

If there is a value x of X for which none of these conditions hold, then unconfoundedness does not hold as well.

8. Another Example

The example displayed in Table 5 serves to illustrate the general procedures to falsify unconfoundedness outlined in the last section. Of course, a simple look at treatment

condition x_2 reveals that the requirements formulated in Definition 1 do not hold in this example. For x_2 we neither have constant assignment probabilities across units nor constant individual conditional expected values $E(Y \mid X = x_2,\ U = u)$ across units. As a consequence, the conditional expected value $E(Y \mid X = x_2)$ is causally biased. Whereas the causally unbiased expected value is $CUE(Y \mid X = x_2) = 105$, the conditional expected value $E(Y \mid X = x_2)$ is equal to 115. Hence, the prima facie effect $PFE(1, 2) = E(Y \mid X = x_1) - E(Y \mid X = x_2) = 115 - 115 = 0$ is considerably different from the average causal effect $ACE(1, 2) = 115 - 105 = 10$. Whereas the average causal effect is positive, the prima facie effect is zero.

In the last paragraph we used Definition 1 to check unconfoundedness. However, in empirical applications, the data on the individual unit level are neither available nor estimable. Hence, Definition 1 cannot be used in empirical falsification trials. What is estimable, however, are the data within subpopulations such as gender (see column 3 of Table 5). The conditional expected values $E(Y \mid X = x,\ W = w)$ may be computed by the formula

$$E(Y \mid X = x,\ W = w) = \sum_u E(Y \mid X = x,\ U = u,\ W = w)\ P(U = u \mid X = x,\ W = w)\ ,$$

where $E(Y \mid X = x,\ W = w,\ U = u) = E(Y \mid X = x,\ U = u)$, because $W = f(U)$. Hence, for this purpose we have to compute the conditional probabilities $P(U = u \mid X = x,\ W = w)$. Using the data of Table 5 we receive for $X = x_2$ and $W = m$:

$$E(Y \mid X = x_2,\ W = m) = 68 \cdot 1/6 + 81 \cdot 1/6 + 89 \cdot 2/6 + 102 \cdot 2/6 = 88.5,$$

and for $X = x_2$ and $W = f$:

$$E(Y \mid X = x_2,\ W = f) = 112 \cdot 3/14 + 119 \cdot 3/14 + 131 \cdot 4/14 + 138 \cdot 4/14 = 126.36.$$

The unconditional probabilities $P(W = w)$ are

$$P(W = f) = P(W = m) = 1/2,$$

whereas the conditional probabilities $P(W = w \mid X = x_2)$ are

$$P(W = m \mid X = x_2) = 3/10 \qquad \text{and} \qquad P(W = f \mid X = x_2) = 7/10.$$

Inserting these results into

$$\sum_w E(Y \mid X = x_2,\ W = w)\ [P(W = w) - P(W = w \mid X = x_2)],$$

[see Eq. (19)] yields

$$88.5\ (1/2 - 3/10) + 126.36\ (1/2 - 7/10) = 17.7 - 25.272 = -7.572.$$

Hence, this is a contradiction to the hypothesis that the regression $E(Y|X)$ is unconfounded in the example of Table 5 [see Eq. (19)].

The second general procedure outlined in section 7 is easier to follow in this example. It easily seen from Table 5 that neither Equation (20) nor Equation (21) holds for $X = x_2$ and $W = m$, for instance.

It should be emphasized again that all parameters necessary for the general procedures to falsify unconfoundedness illustrated in this example are estimable in "real" empirical applications. In this sense unconfoundedness is empirically falsifiable.

Table 5. An example in which the treatment regression $E(Y \mid X)$ is confounded

Observational-unit variable U	$P(U = u)$	W (gender)	$P(X = x_1 \mid U = u)$	$E(Y \mid X = x_1,\ U = u)$	$P(X = x_2 \mid U = u)$	$E(Y \mid X = x_2,\ U = u)$	$P(X = x_3 \mid U = u)$	$E(Y \mid X = x_3,\ U = u)$
u_1	1/8	m	1/2	82	1/10	68	4/10	110
u_2	1/8	m	1/2	89	1/10	81	4/10	110
u_3	1/8	m	1/2	101	2/10	89	3/10	110
u_4	1/8	m	1/2	108	2/10	102	3/10	110
u_5	1/8	f	1/2	118	3/10	112	2/10	110
u_6	1/8	f	1/2	131	3/10	119	2/10	110
u_7	1/8	f	1/2	139	4/10	131	1/10	110
u_8	1/8	f	1/2	152	4/10	138	1/10	110

Note: The (unconditional) probabilities for the three treatments are $P(X = x_1) = 1/2$, $P(X = x_2) = P(X = x_3) = 1/4$.

9. Special Procedures to Falsify Unconfoundedness

In the last two sections, we did not presuppose anything about the functional form of the regression $E(Y|X, W)$. In empirical applications, the procedure outlined in these sections can be used if the sample is big enough to estimate each conditional expected value $E(Y|X = x, W = w)$ sufficiently well and if the probabilities $P(X = x|W = w)$ are known or can also be estimated sufficiently well. If, however,

compared to the sample size, the number of different combinations (x, w) of values of X and W is too large, then there might even be pairs (x, w) for which there are no observations at all. If, however, we are able to make an assumption about the functional form of the regression $E(Y \mid X, W)$ sparseness of data are less of a problem. We will consider two such assumptions: linearity [Eq. (22)] and conditional linearity [Eq. (32)]. The case of linearity has already been considered by Pratt and Schlaifer (1988), for instance.

In the next theorem, $\boldsymbol{X} = (X_1, \ldots, X_p)'$ denotes a p-dimensional regressor and $\boldsymbol{W} = (W_1, \ldots, W_q)'$ a q-dimensional function of U. Both, $\boldsymbol{X}$ and $\boldsymbol{W}$ are column vectors.[15]

Theorem 7. *If (a)* $\boldsymbol{W} = f(U)$*, (b) the treatment regression* $E(Y \mid \boldsymbol{X})$ *is unconfounded (or at least unconfounded with respect to* $\boldsymbol{W}$*), (c) the covariance matrix* $\boldsymbol{\Sigma}$ *of the* $(p + q)$*-dimensional row vector* $(\boldsymbol{X}', \boldsymbol{W}')$ *is regular, and (d)*

$$E(Y \mid \boldsymbol{X}, \boldsymbol{W}) = \beta_0 + \boldsymbol{X}' \boldsymbol{\beta}_X + \boldsymbol{W}' \boldsymbol{\beta}_W, \qquad \beta_0 \in \mathbb{R}, \boldsymbol{\beta}_X \in \mathbb{R}^p, \boldsymbol{\beta}_W \in \mathbb{R}^q, \tag{22}$$

then

$$E(Y \mid \boldsymbol{X}) = \alpha_0 + \boldsymbol{X}' \boldsymbol{\alpha}_X, \qquad \alpha_0 \in \mathbb{R}, \boldsymbol{\alpha}_X \in \mathbb{R}^p \tag{23}$$

$$\boldsymbol{\alpha}_X = \boldsymbol{\beta}_X, \tag{24}$$

$$E(\boldsymbol{W}' \mid \boldsymbol{X})\, \boldsymbol{\beta}_W = E(\boldsymbol{W}')\, \boldsymbol{\beta}_W, \tag{25}$$

$$\boldsymbol{\Sigma}_{XW} \boldsymbol{\beta}_W = \boldsymbol{0}. \tag{26}$$

In Equation (26) $\boldsymbol{\Sigma}_{XW}$ denotes the covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{W}$.

Hence, under the linearity assumption [Eq. (22)], unconfoundedness of $E(Y \mid \boldsymbol{X})$ can be falsified in a specific empirical application, if it is shown that $\boldsymbol{\alpha}_X = \boldsymbol{\beta}_X$ does not hold. Alternatively, one might try to show that $\boldsymbol{\Sigma}_{XW} \boldsymbol{\beta}_W = \boldsymbol{0}$, does not hold. Hence, tests of $\boldsymbol{\alpha}_X = \boldsymbol{\beta}_X$ and tests of $\boldsymbol{\Sigma}_{XW} \boldsymbol{\beta}_W = \boldsymbol{0}$ are *tests of unconfoundedness*, provided that $W = f(U)$ and Equation (22) is true. Usually, a test of linearity or goodness of fit of Equation (22) should precede a test of unconfoundedness. Also note that, given the presuppositions of Theorem 7 [including Eq. (22)], Equations (24) and (26) are equivalent.

Observe that the vectors of regression coefficients $\boldsymbol{\beta}_X$ and $\boldsymbol{\beta}_W$ can be computed by

$$\boldsymbol{\beta}_X = (\boldsymbol{\Sigma}_{XX}^{-1} + \boldsymbol{\Gamma}\, \mathbf{E}^{-1}\, \boldsymbol{\Gamma}')\, \boldsymbol{\Sigma}_{XY} - \boldsymbol{\Gamma}\, \mathbf{E}^{-1} \boldsymbol{\Sigma}_{WY}, \tag{27}$$

[15] Note that all concepts treated in this paper are not restricted to a univariate treatment variable X. If X is p-variate, a value x of X consists of p values $x_1, \ldots, x_p$. Since we are now formulating the models in terms of matrix algebra, we will denote the treatment variable by the boldface letter $\boldsymbol{X}$. The same holds true, of course, for $\boldsymbol{W}$.

and

$$\boldsymbol{\beta}_W = \mathbf{E}^{-1} \, (\boldsymbol{\Sigma}_{WY} \, - \boldsymbol{\Gamma} \, ' \, \boldsymbol{\Sigma}_{XY}), \tag{28}$$

where $\mathbf{E} := \boldsymbol{\Sigma}_{WW} \, - \boldsymbol{\Sigma}_{WX} \, \boldsymbol{\Sigma}_{XX}^{-1} \, \boldsymbol{\Sigma}_{XW}$ and $\boldsymbol{\Gamma} := \boldsymbol{\Sigma}_{XX}^{-1} \, \boldsymbol{\Sigma}_{XW}$ (see, e.g., Seber, 1984). Also note that, for $p = q = 1$, Equations (27) and (28) simplify to the well-known formulas for partial regression coefficients:

$$\beta_X = \left(\sigma_W^2 \, \sigma_{XY} - \sigma_{WX} \, \sigma_{WY} \right) \, / \, \left(\sigma_X^2 \, \sigma_W^2 - \sigma_{WX}^2 \right) \tag{29}$$

and

$$\beta_W = \left(\sigma_X^2 \, \sigma_{WY} - \sigma_{WX} \, \sigma_{XY} \right) \, / \, \left(\sigma_X^2 \, \sigma_W^2 - \sigma_{WX}^2 \right). \tag{30}$$

In Theorem 7 we presuppose that the regression $E(Y \mid \boldsymbol{X}, \, \boldsymbol{W})$ is additive. Especially, Equation (22) does not allow for multiplicative terms such as $X \cdot \, W$, which occur, e.g., in the following equation:

$$\begin{aligned} E(Y \mid X, \, W) \; &= \beta_0 + \beta_1 \, X + \beta_2 \, W + \beta_3 \, X \cdot \, W \\ &= (\beta_0 + \beta_2 \, W) + (\beta_1 + \beta_3 \, W) \, X \\ &= g_0(W) + g_1(W) \, X. \end{aligned} \tag{31}$$

Note that $E(Y \mid X, \, W)$ can always be parameterized in the form of Equation (31), if both X and W are dichotomous. In other cases, however, Equation (31) may not hold for the regression $E(Y \mid X, \, W)$. The values of the *slope coefficient function* $g_1(W)$ are the values of the slope coefficients of the $(W = w)$-conditional linear regressions $E_{W = w}(Y \mid X) = g_0(w) \, + \, g_1(w) \, X$. If, e.g., W is "gender", $E_{W = m}(Y \mid X)$ may represent the regression of Y on X in the subpopulation of males (m).

In the next theorem, we consider a special case which allows for a p-dimensional regressor $\boldsymbol{X} = (X_1, \, ..., \, X_p) \, '$. The basic idea behind Equation (32) is that for each given value w of W, the regression of Y on X is an additive linear multiple regression with the p-variate numerical regressor $\boldsymbol{X} = (X_1, \, ..., \, X_p) \, '$.

Theorem 8. *If (a) the treatment regression $E(Y \mid \boldsymbol{X})$ is unconfounded, (b) X: $\boldsymbol{\Omega} \rightarrow$ IR^p, (c) $\boldsymbol{W} = f (U)$, and (d) there are functions g_0, …, g_p such that*

$$E(Y \mid \boldsymbol{X}, \, \boldsymbol{W}) = g_0(\boldsymbol{W}) + g_1(\boldsymbol{W}) \, X_1 + \, ... \, + g_p(\boldsymbol{W}) \, X_p, \tag{32}$$

then

$$E(Y \mid \boldsymbol{X}) = \alpha_0 + \alpha_1 \, X_1 + \, ... \, + \alpha_p \, X_p \tag{33}$$

with

$$\alpha_0 = E[g_0(\boldsymbol{W})], \quad \alpha_1 = E[g_1(\boldsymbol{W})], \quad \ldots \quad , \alpha_p = E[g_p(\boldsymbol{W})]. \tag{34}$$

According to this theorem, the partial regression coefficients of the regressors X_1, ..., X_p are the expected values of the corresponding $(W = w)$-conditional partial regression coefficients if unconfoundedness holds. Furthermore, this result may also be used for falsification trials of the hypothesis that $E(Y \mid \boldsymbol{X})$ is unconfounded. If the Equations (33) and (34) do not hold, although Equation (32) does, then we can reject the hypothesis that $E(Y \mid \boldsymbol{X})$ is unconfounded. For $p = 1$, this theorem shows how the slope coefficient α_1 of a *simple linear* regression $E(Y \mid X) = \alpha_0 + \alpha_1 X$ is related to the slope coefficient function $g_1(W)$, if unconfoundedness holds. A special case of Equation (32) results if $X_1 := X^1$, ..., $X_p := X^p$, i.e., if the regression of Y on X is a polynomial function of the one-dimensional numeric regressor X.

10. Discussion

In this paper we introduced the concept of unconfoundedness of a treatment regression $E(Y \mid X)$ and showed its relationship to Neyman´s and Rubin´s concepts of individual and average causal effects. We discussed several criteria that are relevant in causal modeling. The weakest is causal unbiasedness of the conditional expected values $E(Y \mid X = x)$ [see Eq. (1)] which has the disadvantage of being nonfalsifiable. Other conditions were independence of U and X, as well as unit-treatment homogeneity: $E(Y \mid X, U) = E(Y \mid X)$ and the and/or-combination of the two. All these conditions are sufficient conditions for causal unbiasedness but they are unnecessarily strong. We finally introduced unconfoundedness as the weakest empirically falsifiable sufficient condition for causal unbiasedness. We defined a regression $E(Y \mid X)$ to be unconfounded if, for each value x of X: (a) $P(X = x \mid U = u) = P(X = x)$ for each value u of U, or (b) $E(Y \mid X = x, U = u) = E(Y \mid X = x)$ for each value u of U. This concept goes beyond the observational level and still has many implications, some of which are empirically testable.

Unconfoundedness implies that the conditional expected values $E(Y \mid X = x)$ and their differences $E(Y \mid X = x_i) - E(Y \mid X = x_j)$, the prima facie effects, are causally unbiased, i.e. $E(Y \mid X = x_i) - E(Y \mid X = x_j) = ACE(i, j)$. This provides the link to the concept of the average causal effect. Furthermore, we showed that unconfoundedness of $E(Y \mid X)$ is sufficient to imply that the conditional expected values $E(Y \mid X = x, W = w)$ and their differences $E(Y \mid X = x_i, W = w) - E(Y \mid X = x_j, W = w)$ are causally unbiased in every subpopulation $W = w$ (see Corollary 2).

The *sufficient conditions* for unconfoundedness of $E(Y|X)$ may guide us in the design of experiments. Most important, however, they make clear that random assignment of units to experimental conditions serves to secure stochastic independence of X and U, which implies unconfoundedness and causal unbiasedness. Such an experiment with random assignment of observational units to experimental conditions is in fact the only way to deliberately create the circumstances in which average causal effects can be estimated. Although unit-treatment homogeneity, [i.e., $E(Y|X, U) = E(Y|X)$] is also sufficient for unconfoundedness (see Th. 3), this condition can not *deliberately* be created through techniques of experimental design. Nevertheless, both sufficient conditions for unconfoundedness, independence of X and U, and unit-treatment homogeneity, show that causal modeling in randomized and nonrandomized experiments aims at the same goal: to secure causal unbiasedness. Hence, if it were possible to generalize the theory presented such that it could also cover causal modeling in observational (i.e., nonexperimental) studies, in which the X-variables do not represent experimental conditions, the two traditions of theorizing about causality in statistics identified by Cox (1992), experimental and observational, could be integrated into a single theory.

The *necessary conditions* for unconfoundedness of $E(Y|X)$ may guide us in nonrandomized experimental causal modeling, because they provide empirically testable consequences. The theorems presented show which statistical hypotheses have to be tested and rejected in order to falsify the hypothesis of unconfoundedness of $E(Y|X)$ in a specific application.

What happens if a $\boldsymbol{W}$ has been identified with respect to which there is confounding? First, we may look at the regressive dependencies of Y on $\boldsymbol{X}$ *within each subpopulation* defined by $\boldsymbol{W} = \boldsymbol{w}$, i.e., we may look at the *conditional effects*. In fact, this provides more detailed information about the conditional regressive dependence of Y on $\boldsymbol{X}$ given $\boldsymbol{W} = \boldsymbol{w}$. Note that this procedure is in full accordance with the framework presented in this paper. One would just replace the total population $\boldsymbol{\Omega}_U$ by the subpopulation defined by $\boldsymbol{W} = \boldsymbol{w}$. This procedure is, in principle, equivalent with statistically controlling for the covariate vector $\boldsymbol{W}$ provided that the equation for the regression $E(Y|X, W)$ is adequately specified. Second, if we are willing to assume that the conditional regressions $E_{W=w}(Y|X)$ are unbiased, we may also compute the average causal effects via Equation (16), because, under this assumption

$$ACE(i, j) = E_{adj\ for\ W}(Y|X = x_i) - E_{adj\ for\ W}(Y|X = x_j)$$

(for details, see Wüthrich-Martone, Steyer, Nachtigall & Suhl, 1999).

A limitation requiring a more general formulation of our approach follows from assuming $P(X = x, U = u) > 0$. Whereas this assumption has been useful for a simple presentation, it restricts generality because it is not compatible with the assumption

that X has a normal distribution, for instance, which is often made in nonexperimental causal modeling (see, e.g., Bollen, 1989). Since normality is not really necessary in structural equation models (see, e.g., Browne & Arminger, 1995), this restriction is not too serious. Nevertheless, we should look for a more general theory which is not restricted to discrete treatments and units.

Another open problem is how the theory presented relates to statistical sampling models. Especially, statistical tests and procedures have to be developed to help us in deciding whether or not unconfoundedness holds in a specific application. For some cases, these tests and procedures have already been presented (see, e.g., Allison, 1995; Clogg et al., 1992; Clogg et al., 1995; Steyer, Gabler, Rucai & Schuster, 1997; von Davier, 2000), for others they still need to be developed.

A significance test of unconfoundedness in very large samples may not really be meaningful in many empirical applications, because the exact null hypothesis will rarely hold in nonrandomized experiments. Instead, one might rather be interested in estimating a coefficient reflecting the strength of confounding of a regression $E(Y|X)$ with respect to a specific variable W. Such a coefficient has been proposed by Steyer, Gabler, and Rucai (1996). More research and experience will be necessary to work out principles dealing with the practical relevance of confoundings in concrete empirical studies.

References

[1] Allison, P. D. (1995). The impact of random predictors on comparisons of coefficients between models: Comment on Clogg, Petkova, and Haritou. *American Journal of Sociology*, *100*, 1294-1305.

[2] Bollen, K. A. (1989). *Structural equation models with latent variables.* New York: Wiley.

[3] Browne, M. W. & Arminger, G. (1995). Specification and estimation of mean- and covariance-structure models. In G. Arminger, C. C. Clogg & M. E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 185-250). New York: Plenum.

[4] Clogg, C. C., Petkova, E., & Shidadeh, E. S. (1992). Statistical methods for analyzing collapsibility in regression models. *Journal of Educational Statistics*, *17*, 51-74.

[5] Clogg, C. C., Petkova, E., & Haritou, A. (1995). Statistical methods for comparing regression coefficients between models. *American Journal of Sociology*, *100*, 1261-1293.

[6] Cook, T. D. & Campbell, D. T. (1979). *Quasi-experimentation: design and analysis issues for field settings*. Boston: Houghton Mifflin.

[7] Cox, D. R. (1992). Causality: Some statistical aspects. *Journal of the Royal Statistical Society, Series A*, *155*, 291-301.

[8] Holland, P. (1986). Statistics and causal inference (with comments). *Journal of the American Statistical Association*, *81*, 945-970.

[9] Holland, P. W. (1988a). Causal inference in retrospective studies. *Evaluation Review*, *13*, 203-231.

[10] Holland, P. W. (1988b). Causal inference, path analysis, and recursive structural equations models. *Sociological Methodology*, *18*, 449-484.

[11] Holland, P. W. & Rubin, D. B. (1983). On Lord's paradox. In H. Wainer & S. Messick (eds.), *Principals of modern psychological measurement* (pp. 3-25). Hillsdale, NJ: Erlbaum.

[12] Neyman, J. (with Iwaszkiewicz, K., and Kolodziejczyk, S.). (1935). Statistical problems in agricultural experimentation (with discussion). *Supplement of Journal of the Royal Statistical Society*, *2*, 107-180.

[13] Neyman, J. (1923/1990). On the application of probability theory to agricultural experiments. Essay on principles. Section 9. *Statistical Science*, *5*, 465-472.

[14] Pearl, J. (1998). Why there is no statistical test for confounding, why many think there is, and why they are almost right. (Technical Report R-256, January 1998).

[15] Pratt, J. W. & Schlaifer, R. (1988). On the interpretation and observation of laws. *Journal of Econometrics*, *39*, 23-52.

[16] Rosenbaum, P. R. (1984a). Conditional permutation tests and the propensity score in observational studies. *Journal of the American Statistical Association*, *79*, 565-574.

[17] Rosenbaum, P. R. (1984b). The consequences of adjustment for a concomitant variable that has been affected by the treatment. *Journal of the Royal Statistical Society, Series A*, *147*, 656-666.

[18] Rosenbaum, P. R. (1984c). From association to causation in observational studies: The role of tests of strongly ignorable treatment assignment. *Journal of the American Statistical Association*, *79*, 41-48.

[19] Rosenbaum, P. R. & Rubin, D. B. (1983a). Assessing sensitivity to an unobserved binary covariate in an observational study with binary outcome. *Journal of the Royal Statistical Society, Series B*, *45*, 212-218.

[20] Rosenbaum, P. R. & Rubin, D. B. (1983b). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*, 41-55.

[21] Rosenbaum, P. R. & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*, 516-524.

[22] Rosenbaum, P. R. & Rubin, D. B. (1985a). The bias due to incomplete matching. *Biometrics*, *41*, 103-116.

[23] Rosenbaum, P. R. & Rubin, D. B. (1985b). Constructing a control group using multivariate matched sampling methods that incorporate the propensity score. *The American Statistician*, *39*, 33-38.

[24] Rubin, D. (1973a). The use of matched sampling and regression adjustment to remove bias in observational studies. *Biometrics*, *29*, 185-203.

[25] Rubin, D. B. (1973b). Matching to remove bias in observational studies. *Biometrics*, *29*, 159-183.

[26] Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*, 688-701.

[27] Rubin, D. B. (1977). Assignment of treatment group on the basis of a covariate. *Journal of Educational Statistics*, *2*, 1-26.

[28] Rubin, D. B. (1978). Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, *6*, 34-58.

[29] Rubin, D. B. (1985). The use of propensity scores in applied Bayesian inference. *Bayesian Statistics*, *2*, 463-472.

[30] Rubin, D. B. (1986). Which ifs have causal answers. *Journal of the American Statistical Association*, *81*, 961-962.

[31] Rubin, D. B. (1990). Comment: Neyman (1923) and causal inference in experiments and observational studies. *Statistical Science*, *5*, 472-480.

[32] Seber, G. A. F. (1984). *Multivariate observations.* New York: Wiley.

[33] Sobel, M. E. (1994). Causal inference in latent variables analysis. In A. von Eye & C. C. Clogg (eds.), *Latent variables analysis* (pp. 3-35). Thousand Oaks, CA: Sage.

[34] Sobel, M. E. (1995). Causal inference in the Social and Behavioral Sciences. In G. Arminger, C. C. Clogg & M. E. Sobel (eds.), *Handbook of Statistical Modeling for the Social and Behavioral Sciences* (pp. 1-38). New York: Plenum.

[35] Steyer, R. (1992). *Theorie kausaler Regressionsmodelle* [Theory of causal regression models]. Stuttgart: Gustav Fischer.

[36] Steyer, R., Gabler, S., von Davier, A. A., Nachtigall, C. & Buhl. T. (2000). Theory of causal regression models: Individual and average causal effects in regression models with discrete units and discrete treatment variable. *Methods of Psychological Research-Online*,

[37] Steyer, R. Gabler, S. & Rucai, A. A. (1996). Individual Causal Effects, Average Causal Effects, and Unconfoundedness in Regression Models. In F. Faulbaum and W. Bandilla (Eds.), *SoftStat '95. Advances in Statistical Software 5* (pp. 203-210). Stuttgart: Lucius & Lucius.

[38] Steyer, R., von Davier, A. A., Gabler, S. & Schuster, C. (1997). Testing unconfoundedness in linear regression models with stochastic regressors. In W. Bandilla & F. Faulbaum (Eds.), *SoftStat '97. Advances in Statistical Software 6* (pp. 377-384). Stuttgart: Lucius & Lucius.

[39] Steyer, R. & Eid, M. (2001). *Messen und Testen.* [Measuring and testing. 2nd edition]. Berlin: Springer.

[40] Von Davier, A. A. (2000). *Tests of unconfoundedness in regression models with normally distributed variables.* Unpublished doctoral dissertation, University of Magdeburg, Germany.

[41] Williams, D. (1991). *Probability with martingales.* Cambridge: Cambridge University Press.

[42] Wüthrich-Martone, O., Steyer, R., Nachtigall, C. & Suhl, U. (1999). Causality, confounding and unbalanced analysis of variance. In: H. Friedl, A. Berghold and G. Kauermann (eds.), Statistical Modelling. Proceedings of the 14th International Workshop on Statistical Modelling in Graz, Austria, July 19-23, 1999 (pp. 719-722).

Appendix A: Extended Notes

Note 1. Strong ignorability implies causal unbiasedness.

$$
\begin{aligned}
E(Y \mid X = x_i) &= \sum\nolimits_u E(Y \mid X = x_i,\, U = u)\; P(U = u \mid X = x_i) && \text{[always true if } U \text{ is discrete]} \\
&= \sum\nolimits_u Y_i(u)\; P(U = u \mid X = x_i) && \text{[definition of } Y_i\text{]} \\
&= \sum\nolimits_u Y_i(u)\; P[Y_i = Y_i(u) \mid X = x_i] && \text{[see comment below]} \\
&= E(Y_i \mid X = x_i) && \text{[definition of conditonal expected value]} \\
&= E(Y_i) && \text{[because of strong ignorability]}
\end{aligned}
$$

$$= \sum_u \; Y_i(u) \; P[Y_i = Y_i(u)] \qquad \text{[definition of expected value]}$$

$$= \sum_u \; Y_i(u) \; P(U = u) \qquad \text{[see comment below]}$$

$$= \sum_u \; E(Y \mid X = x_i,\, U = u) \; P(U = u) \qquad \text{[definition of } Y_i\text{]}$$

$$= CUE(Y \mid X = x_i). \qquad \text{[definition of the causally unbiased expected value]}$$

Comment. Note that the independence condition $Y_1, \ldots, Y_{n_x} \perp X$ in strong ignorability implies regressive independence $E(Y_i \mid X = x) = E(Y_i)$ for each pair $(i,\, x)$ of indices $i \in \{1, \ldots, n_x\}$ and values x of X and that the events $\{Y_i = Y_i(u)\} := \{\omega \in \Omega\colon Y_i(\omega) \in \{Y_i(u)\}\}$ and $\{U = u\} := \{\omega \in \Omega\colon U(\omega) \in \{u\}\}$ are identical subsets of Ω and therefore have identical conditional and unconditional probabilities. This does not presume that all the variables Y_i or one them are one-to-one functions of U. It is still possible that two different units u_1 and u_2 have identical values $Y_i(u_1) = Y_i(u_2)$.

Appendix B: Proofs

Proof of Theorem 1. In this proof we will use the so-called *tower property* of regressions (see, e.g., Williams, 1991, p. 88 or Steyer & Eid, 2001, p. 357):

$$E\,[Y \mid f\,(X)] = E\,[E(Y \mid X\,) \mid f\,(X)], \tag{35}$$

and its special case

$$E\,(Y\,) = E\,[E(Y \mid X\,)] \tag{36}$$

for $f\,(X)$ = constant. Equation (36) is always true and Equation (35) is always true for every (measurable) mapping $f\,(X)$ of X. If unconfoundedness as defined in Definition 1 holds, then, for each value x of X, Equation (3) or Equation (4) will hold. Let us assume that Equation (4) holds. This equation may also be written

$$E_{X=x}(Y \mid U\,) = E_{X=x}(Y), \tag{37}$$

where the $E_{X=x}$ means that we refer to the expectation (or conditional expection) with respect to the conditional probability measure $P_{X=x}$. Since W is a measurable mapping of U we may apply Equation (35) which yields:

$$E_{X=x}(Y \mid W\,) \;= E_{X=x}[E_{X=x}(Y \mid U\,) \mid W\,] \qquad \text{[see Eq. (35)]}$$

$$= E_{X=x}[E_{X=x}(Y\,) \mid W\,] = E_{X=x}(Y). \qquad \text{[see Eqs. (37) and (36)]}$$

The corresponding argument may be applied to the case in which Equation (3) holds. As $P(X = x \mid U = u) = E(I_{X=x} \mid U = u)$, where $I_{X=x}$ denotes the indicator of the event $X = x$, Equation (3) may also be written

$$E(I_{X=x} \mid U) = E(I_{X=x}). \tag{38}$$

Since W is a measurable mapping of U we may again apply Equation (35) which, together with Equation (38) yields:

$$E(I_{X=x} \mid W) = E[E(I_{X=x} \mid U) \mid W] \qquad \text{[see Eq. (35)]}$$

$$= E[E(I_{X=x}) \mid W] = E(I_{X=x}). \qquad \text{[see Eqs. (38) and (36)]}$$

However, this is just another way to write Equation (7).

The other direction of implication is more simple to prove: The observational unit variable U itself may be written $f(U)$, because $U = id(U)$, where id is the identity mapping.

The following lemma will be needed in the proof of Theorem 2.

Lemma 1. *If Equation* (9) *holds for each* $W = f(U)$*, then, for each value* x *of* X *and each value* w *of* W*,*

$$P(W = w \mid X = x) = P(W = w) \quad \textit{or} \quad E(Y \mid X = x,\ W = w) = E(Y \mid X = x). \tag{39}$$

Proof of Lemma 1. Let $I_{W=w}$ denote the indicator variable for the event $W = w := \{\omega \in \Omega\colon W(\omega) = w\}$ and note that $I_{W=w} = f(U)$. Hence, Equation (9) implies:

$$E(Y \mid X = x) = E(Y \mid X = x,\ I_{W=w} = 1)\ P(W = w) + E(Y \mid X = x,\ I_{W=w} = 0)\ P(W \neq w). \tag{40}$$

Another equation for $E(Y \mid X = x)$ which is always true [we still assume $P(X = x,\ U = u) > 0$ implying $P(X = x,\ W = w) > 0$)] is:

$$E(Y \mid X = x) = E(Y \mid X = x,\ I_{W=w} = 1)\ P(W = w \mid X = x) + E(Y \mid X = x,\ I_{W=w} = 0)\ P(W \neq w \mid X = x). \tag{41}$$

If we subtract Equation (41) from (40), we receive:

$$E(Y \mid X = x,\ I_{W=w} = 1)\ [P(W = w) - P(W = w \mid X = x)]$$

$$= E(Y \mid X = x,\ I_{W=w} = 0)\ [P(W = w) - P(W = w \mid X = x)].$$

This implies $P(W = w) = P(W = w \mid X = x)$ or $E(Y \mid X = x,\ I_{W=w} = 1) = E(Y \mid X = x,\ I_{W=w} = 0)$, and $P(X = x) = P(X = x \mid W = w)$ or $E(Y \mid X = x,\ W = w) = E(Y \mid X = x)$, which was to be shown.

Proof of Theorem 2. If unconfoundedness as defined in Definition 1 holds, then Equation (7) or (8) will be true (see Th. 1). We first show that Equation (7) as well as Equation (8) imply Equation (9).

If Equation (7) holds, then

$$P(W = w \mid X = x) = P(W = w) \qquad \textit{for each value } w \textit{ of } W \tag{42}$$

will hold as well. [We still presume $P(X = x,\ U = u) > 0$, which implies $P(X = x,\ W = w) > 0$, $P(X = x) > 0$, and $P(W = w) > 0$].

Hence, using Equation (11) – which is always true – and (42), Equation (9) follows. If Equation (8) holds, then

$$\sum_w E(Y \mid X = x,\ W = w)\ P(W = w) = \sum_w E(Y \mid X = x)\ P(W = w)$$

$$= E(Y \mid X = x) \sum_w P(W = w) = E(Y \mid X = x), \qquad \text{for each value } x \text{ of } X,$$

which again is Equation (9).

The other direction of the implication is more difficult to prove. We will conduct an indirect proof: Assuming neither Equation (7) nor Equation (8) were true will lead to a contradiction which together with Theorem 1 completes the proof.

Assume, for each $W = f(U)$, Equation (9) holds and there were a $W = f(U)$ with two values w_1, w_2 such that

$$P(W = w_1 \mid X = x) \neq P(W = w_1) \quad \text{and} \quad E(Y \mid X = x,\ W = w_2) \neq E(Y \mid X = x), \tag{43}$$

for the same value x of X. For $w_1 = w_2$, this would contradict Lemma 1. Hence, according to Lemma 1:

$$P(W = w_1 \mid X = x) \neq P(W = w_1) \quad \text{and} \quad E(Y \mid X = x,\ W = w_1) = E(Y \mid X = x), \tag{44}$$

$$E(Y \mid X = x,\ W = w_2) \neq E(Y \mid X = x) \quad \text{and} \quad P(W = w_2 \mid X = x) = P(W = w_2), \tag{45}$$

as well as $w_1 \neq w_2$. This is tantamount to assuming that neither Equation (7) nor Equation (8) were true. We now show that this leads to a contradiction. Let I_{12} be an indicator variable such that $I_{12} = 1$ if $W = w_1$ or $W = w_2$ and $I_{12} = 0$ otherwise. Then

$$P(I_{12} = 1 \mid X = x) = P(W = w_1 \mid X = x) + P(W = w_2 \mid X = x) = P(W = w_1 \mid X = x) + P(W = w_2).$$

If $P(I_{12} = 1 \mid X = x) = P(I_{12} = 1)$, then $P(I_{12} = 1 \mid X = x) = P(W = w_1) + P(W = w_2)$. But this implies $P(W = w_1 \mid X = x) = P(W = w_1)$, which would contradict (44). Hence,

$$P(I_{12} = 1 \mid X = x) \neq P(I_{12} = 1) \quad \text{and} \quad E(Y \mid X = x) = E(Y \mid X = x,\ I_{12} = 1), \tag{46}$$

again using Lemma 1. We have

$$E(Y \mid X = x,\ I_{12} = 1) = E(Y \mid X = x,\ W = w_1)\ P(W = w_1 \mid X = x)\ /\ P(I_{12} = 1 \mid X = x)$$

$$+\ E(Y \mid X = x,\ W = w_2)\ P(W = w_2 \mid X = x)\ /\ P(I_{12} = 1 \mid X = x). \tag{47}$$

Inserting (46) as well as (44) into (47) yields

$$E(Y \mid X = x) - E(Y \mid X = x)\ P(W = w_1 \mid X = x)\ /\ P(I_{12} = 1 \mid X = x)$$

$$= E(Y \mid X = x,\ W = w_2)\ P(W = w_2 \mid X = x)\ /\ P(I_{12} = 1 \mid X = x).$$

Now

$$P(W = w_1 \mid X = x) \;/\; P(I_{12} = 1 \mid X = x) = 1 - P(W = w_2 \mid X = x) \;/\; P(I_{12} = 1 \mid X = x)$$

implies

$$[E(Y \mid X = x) - E(Y \mid X = x,\ W = w_2)]\ P(W = w_2 \mid X = x) \;/\; P(I_{12} = 1 \mid X = x) = 0.$$

It follows that $E(Y \mid X = x) = E(Y \mid X = x,\ W = w_2)$ which is a contradiction to (45). Hence, we have shown that if Equation (9) holds for each $W = f(U)$, then Equation (7) or Equation (8) must hold for each $W = f(U)$. Theorem 1 completes the proof.

Proof of Corollary 1. This proposition immediately follows from Theorem 2 and

$$PFE(i, j) = \sum\nolimits_w\ [E(Y \mid X = x_i,\ W = w) - E(Y \mid X = x_j,\ W = w)]\ P(W = w) =$$

$$\sum\nolimits_w\ E(Y \mid X = x_i,\ W = w)\ P(W = w)\ -\ \sum\nolimits_w\ E(Y \mid X = x_j,\ W = w)]\ P(W = w).$$

Proof of Theorem 3. Propositions (i) and (ii) are immediate consequences of Definition 1. Proposition (iii) can be derived as follows: According to Theorem 2 we have to show that Equation (9) holds for each $W = f(U)$.

$$\begin{aligned}
E(Y \mid X = x) &= E(Y \mid X = x) \cdot 1 = E(Y \mid X = x) \cdot \sum\nolimits_w\ P(W = w) \\
&= \sum\nolimits_w\ E(Y \mid X = x) \cdot\ P(W = w) \\
&= \sum\nolimits_w\ [E(Y \mid X = x,\ W = w) - h(w)] \cdot P(W = w) && \text{[see Eq. (12)]} \\
&= \sum\nolimits_w\ E(Y \mid X = x,\ W = w) \cdot P(W = w) - \sum\nolimits_w\ h(w) \cdot P(W = w) \\
&= \sum\nolimits_w\ E(Y \mid X = x,\ W = w) \cdot P(W = w).
\end{aligned}$$

The last equation follows from $\sum_w\ h(w) \cdot P(W = w) = E[h(W)] = 0$, which is true because: $E(Y) = E\,[E(Y \mid X,\ W)] = E\,[E(Y \mid X) + h(W)] = E(Y) + E[h\,(W)]$.

Proof of Theorem 4. (i) The variables Y_i are defined in such a way that they are measurable functions of U. Hence, independence of U implies independence of $Y_1, \ldots, Y_{n_x}$ and X.

(ii) Unit-treatment homogeneity implies that each Y_i is a constant, and constants are always stochastic independent of any random variable.

Proof of Theorem 5. Since U and X are assumed to be discrete, the following equation is always true for each value x of X:

$$E(Y \mid X = x) = \sum\nolimits_u\ E(Y \mid X = x,\ U = u)\ P(U = u \mid X = x). \tag{48}$$

If Equation (3) holds for a given value x of X, then $P(U = u \mid X = x) = P(U = u)$, for each value u of U. Inserting $P(U = u)$ in the last equation yields unbiasedness of each conditional expected value $E(Y \mid X = x)$. If Equation (4) holds, for this value x of X, then

$$\begin{aligned} E(Y \mid X = x) &= E(Y \mid X = x) \sum\nolimits_u P(U = u) \\ &= \sum\nolimits_u E(Y \mid X = x)\, P(U = u) \\ &= \sum\nolimits_u E(Y \mid X = x,\, U = u)\, P(U = u). \end{aligned}$$

This proves proposition (i).

Proposition (ii) now follows from (i) and

$$\begin{aligned} ACE(i, j) &= \sum\nolimits_u [E(Y \mid X = x_i,\, U = u) - E(Y \mid X = x_j,\, U = u)]\, P(U = u). \\ &= \sum\nolimits_u E(Y \mid X = x_i,\, U = u)\; P(U = u) \; - \sum\nolimits_u E(Y \mid X = x_j,\, U = u)\, P(U = u) \\ &= CUE(Y \mid X = x_i) - CUE(Y \mid X = x_j)\ . \end{aligned}$$

Proof of Theorem 6. Let $\Omega_{U_w} := \{u \in \mathbf{\Omega}_U \colon W(\omega) = w\}$, denote the subpopulation represented by $W = w$ and suppose Equation (3) holds. Because of

$$P_{W=w}(X = x \mid U = u) = \begin{cases} P(X = x \mid U = u), & \text{if } u \in \Omega_{U_w} \\ \text{undefined}, & \text{otherwise}, \end{cases}$$

Equation (3) implies

$$P_{W=w}(X = x \mid U = u) = \begin{cases} P(X = x) = P_{W=w}(X = x), & \text{if } u \in \Omega_{U_w} \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

Now suppose Equation (4) holds. Then, because of

$$E_{W=w}(Y \mid X = x, U = u) = \begin{cases} E(Y \mid X = x, U = u), & \text{if } u \in \Omega_{U_w} \\ \text{undefined}, & \text{otherwise}, \end{cases}$$

Equation (4) implies

$$E_{W=w}(Y \mid X = x, U = u) = \begin{cases} E(Y \mid X = x) = E_{W=w}(Y \mid X = x), & \text{if } u \in \Omega_{U_w} \\ \text{undefined}, & \text{otherwise.} \end{cases}$$

This completes the proof, because we have shown that for each value x of X

$$P_{W=w}(X = x \mid U = u) = P_{W=w}(X = x) \quad \text{for each value } u \text{ of } \Omega_{U_w} \tag{49}$$

or

$$E_{W=w}(Y \mid X = x, U = u) = E_{W=w}(Y \mid X = x) \quad \text{for each value } u \text{ of } \Omega_{U_w}\ . \tag{50}$$

Proof of Corollary 2. This is an immediate consequence of Theorem 6 and Theorem 5 (i).

Proof of Theorem 7. For each value $\boldsymbol{x}$ of $\boldsymbol{X}$:

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = \sum_{\boldsymbol{w}} E(Y|\boldsymbol{X}=\boldsymbol{x},\ \boldsymbol{W}=\boldsymbol{w}) \cdot P(\boldsymbol{W}=\boldsymbol{w}) \qquad \text{[see Eq. (9)]}$$

$$= \sum_{\boldsymbol{w}} (\beta_0 + \boldsymbol{x}' \boldsymbol{\beta}_X + \boldsymbol{w}' \boldsymbol{\beta}_W) \cdot P(\boldsymbol{W}=\boldsymbol{w}) \qquad \text{[see Eq. (22)]}$$

$$= \beta_0 + \boldsymbol{x}' \boldsymbol{\beta}_X + E(\boldsymbol{W}') \boldsymbol{\beta}_W$$

$$= [\beta_0 + E(\boldsymbol{W}') \boldsymbol{\beta}_W] + \boldsymbol{x}' \boldsymbol{\beta}_X.$$

Hence, the constant α_0 of the regression $E(Y|\boldsymbol{X})$ is identical with $\beta_0 + E(\boldsymbol{W}') \boldsymbol{\beta}_W$ and the vector of regression coefficients $\boldsymbol{\alpha}_X$ is equal to $\boldsymbol{\beta}_X$, which proves equations (23) and (24).

Equation (25) may be derived in the same way, only using

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = \sum_{\boldsymbol{w}} E(Y|\boldsymbol{X}=\boldsymbol{x},\ \boldsymbol{W}=\boldsymbol{w}) \cdot P(\boldsymbol{W}=\boldsymbol{w}|\boldsymbol{X}=\boldsymbol{x}) \qquad \text{[see Eq. (11)]}$$

to start with. This yields

$$E(Y|\boldsymbol{X}=\boldsymbol{x}) = [\beta_0 + E(\boldsymbol{W}'|\boldsymbol{X}=\boldsymbol{x}) \boldsymbol{\beta}_W] + \boldsymbol{x}' \boldsymbol{\beta}_X.$$

The last equation shows that $E(Y|\boldsymbol{X})$ will be of the form $\alpha_0 + \boldsymbol{X}' \boldsymbol{\alpha}_X$ (see Eq. (23)) with $\boldsymbol{\alpha}_X = \boldsymbol{\beta}_X$ (see Eq. (24)), if $E(\boldsymbol{W}'|\boldsymbol{X}=\boldsymbol{x}) \boldsymbol{\beta}_W = E(\boldsymbol{W}') \boldsymbol{\beta}_W$. This proves Equation (25).

Let $\boldsymbol{\Sigma}_{XW} = Cov(\boldsymbol{X},\ \boldsymbol{W})$ denote the $p \times q$ covariance matrix of $\boldsymbol{X}$ and $\boldsymbol{W}$. Since $Cov(\boldsymbol{X},\boldsymbol{W}) = Cov[\boldsymbol{X}, E(\boldsymbol{W}|\boldsymbol{X})]$,

$$\boldsymbol{\Sigma}_{XW} \boldsymbol{\beta}_W = Cov[\boldsymbol{X}, E(\boldsymbol{W}|\boldsymbol{X})] \ \boldsymbol{\beta}_W$$

$$= Cov[\boldsymbol{X}, E(\boldsymbol{W}'|\boldsymbol{X}) \ \boldsymbol{\beta}_W] \qquad \text{(bilinearity of the covariance)}$$

$$= Cov[\boldsymbol{X}, E(\boldsymbol{W}') \ \boldsymbol{\beta}_W] \qquad \text{[see Eq. (25)]}$$

$$= \boldsymbol{0},$$

because $E(\boldsymbol{W}') \boldsymbol{\beta}_W$ is a constant. This proves Equation (26).

Using $Cov(\boldsymbol{X},\ Y) = Cov[\boldsymbol{X}, E(Y|\boldsymbol{X},\ \boldsymbol{W})]$, $Cov(\boldsymbol{W},\ Y) = Cov[\boldsymbol{W}, E(Y|\boldsymbol{X},\ \boldsymbol{W})]$, and Equation (22) yields:

$$\boldsymbol{\Sigma}_{XY} = \boldsymbol{\Sigma}_{XX} \boldsymbol{\beta}_X + \boldsymbol{\Sigma}_{XW} \boldsymbol{\beta}_W$$

$$\boldsymbol{\Sigma}_{WY} = \boldsymbol{\Sigma}_{WX} \boldsymbol{\beta}_X + \boldsymbol{\Sigma}_{WW} \boldsymbol{\beta}_W.$$

Writing these two equations in a single matrix equation, the well-known results for the inverse of a partitioned matrix (e.g., Seber, 1984, p. 519) yield equations (27) and (28).

Proof of Theorem 8. For each value $\boldsymbol{x}$ of $\boldsymbol{X}$:

$$
\begin{aligned}
E(Y \mid \boldsymbol{X} = \boldsymbol{x}) &= \sum\nolimits_{w} E(Y \mid \boldsymbol{X} = \boldsymbol{x},\ \boldsymbol{W} = \boldsymbol{w}) \cdot P(\boldsymbol{W} = \boldsymbol{w}) && \text{[see Eq. (9)]} \\
&= \sum\nolimits_{w} [g_0(\boldsymbol{w}) + g_1(\boldsymbol{w})\ x_1 + \ldots + g_p(\boldsymbol{w})\ x_p] \cdot P(\boldsymbol{W} = \boldsymbol{w}) && \text{[see Eq. (32)]} \\
&= E[g_0(\boldsymbol{W})] + E[g_1(\boldsymbol{W})]\ x_1 + \ldots + E[g_p(\boldsymbol{W})]\ x_p.
\end{aligned}
$$