Background: Instrumental variables (IVs) can be used to provide evidence as to whether a treatment $X$ has a causal effect on an outcome $Y$. Even if the instrument $Z$ satisfies the three core IV assumptions of relevance, independence, and exclusion restriction, further assumptions are required to identify the average causal effect (ACE) of $X$ on $Y$. Sufficient assumptions for this include homogeneity in the causal effect of $X$ on $Y$; homogeneity in the association of $Z$ with $X$; and no effect modification.

Methods: We describe the no simultaneous heterogeneity assumption, which requires the heterogeneity in the $X$-$Y$ causal effect to be mean independent of (i.e., uncorrelated with) both $Z$ and heterogeneity in the $Z$-$X$ association. This happens, for example, if there are no common modifiers of the $X$-$Y$ effect and the $Z$-$X$ association, and the $X$-$Y$ effect is additive linear. We illustrate the assumption of no simultaneous heterogeneity using simulations and by re-examining selected published studies.

Results: Under no simultaneous heterogeneity, the Wald estimand equals the ACE even if both homogeneity assumptions and no effect modification (which we demonstrate to be special cases of— and therefore stronger than—no simultaneous heterogeneity) are violated.

Conclusions: The assumption of no simultaneous heterogeneity is sufficient for identifying the ACE using IVs. Since this assumption is weaker than existing assumptions for ACE identification, doing so may be more plausible than previously anticipated.

Key words: Causal inference; Effect modification; Homogeneity; Identification; Instrumental variables.
the ACE is identified if no unmeasured confounders are additive modifiers of the association between the instrument and the treatment or of the effect of the treatment on the outcome. Cui and Tchetgen Tchetgen,10 again focusing on binary instrument and treatment, proposed a weaker version of this condition which only requires that there is no unmeasured common additive modifier of the instrument-treatment association and the treatment-outcome effect. Syrgkanis et al.13 states that the ACE is not generally identified if individual-level instrument-treatment and treatment-outcome effects are dependent.13

Numerous reviews and methodologic papers have described several identifying assumptions.4,14-17 Here, we introduce the no simultaneous heterogeneity assumption. We show that if this assumption holds, the IV estimand equals the ACE. We also show that other IV assumptions are special cases of this assumption. Finally, we use simulations to corroborate the theory.

METHODS

Notation and Assumptions

Let \( z, x, y, \) and \( U \) respectively denote the instrument, the time-fixed treatment, the outcome, and all unmeasured common causes of \( X \) and \( Y \). For simplicity, we consider the case where no adjustment is made for measured covariates. However, the concepts developed here can be trivially extended to accommodate measured covariates. We also discuss covariate adjustment in the simulation study (see eAppendix; http://links.lww.com/EDE/C13).

\( Z \) is a valid IV if the following three causal assumptions (illustrated in Figure 1) are satisfied: (1) relevance: \( Z \not\perp \perp U \) (where “\( \not\perp \perp \)” denotes statistical independence); (2) independence: \( Z \perp \perp Y | X, U \); and (3) exclusion restriction: \( Z \not\perp \perp Y | X, U \).

Figure 1 is a graphical representation of the following nonparametric structural equation model:

\[
X_i = f_X(Z = z_i, U_i, \varepsilon_X = \varepsilon_{X_i}) \quad Y_i = f_Y(X = x_i, U_i, \varepsilon_Y = \varepsilon_{Y_i})
\]

where \( X_i \) is the value of \( X \) for individual \( i \) (the same notation applies to other variables), \( f_X \) and \( f_Y \) respectively denote the functions governing \( X \) and \( Y \), and \( \varepsilon_X \) and \( \varepsilon_Y \) respectively denote stochastic direct causes of \( X \) and \( Y \) (so that \( \varepsilon_X \perp \varepsilon_Y \)). We now define \( F_{X_i}(z) = E[X_i | do(Z = z), U = U_i, \varepsilon_X = \varepsilon_{X_i}] \) and \( F_{Y_i}(x) = E[Y_i | do(X = x), U = U_i, \varepsilon_Y = \varepsilon_{Y_i}] \) that is, \( F_{X_i}(z) \) is the expectation of \( X \) when \( Z \) is set (possibly counterfactually) to \( z \), while the other variables retain their observed values. A similar interpretation holds for \( F_{Y_i}(x) \).

The individual-level instrumental and treatment effects \( \beta_X \) and \( \beta_Y \) can be defined as follows:

\[
\beta_{X_i} = F_{X_i}(1) - F_{X_i}(0) \quad \text{for a binary } Z,
\]

\[
\beta_{X_i} = F_{X_i}(z) - F_{X_i}(z - 1) \quad \text{for a multivalued discrete } Z,
\]

\[
\beta_{Y_i} = \frac{\partial}{\partial z} F_{X_i}(z) \Bigg|_{z=z_i} \quad \text{for a continuous } Z,
\]

\[
\beta_{Y_i} = \frac{\partial}{\partial x} F_{Y_i}(x) \Bigg|_{x=x_i} \quad \text{fora continuous } X.
\]

For a multivalued discrete \( Z \), we assume \( Z \) is coded numerically such that \( E[X|Z = 1] \leq \ldots \leq E[X|Z = K] \), where \( K \) is the number of values that \( Z \) attains, and \( \varepsilon \in \{2, \ldots, K\} \). Notice that, for continuous \( Z \), the definition of \( \beta_X \) implicitly assumes that \( F_X(z) \) is differentiable with respect to \( Z \). For noncontinuous \( X \), this would happen for example if \( X_i \sim Bernoulli(p_i) \) or \( X_i \sim Poisson(\lambda_i) \), where \( p_i \) or \( \lambda_i \) are differentiable functions of \( Z \). A similar notion applies to \( \beta_Y \) and \( F_Y(z) \) for a continuous \( X \).

Under the stable unit treatment value assumption, \( \beta_{X_i} \) is the additive change in the expectation of \( X \) caused by a unit increase in \( Z \) in individual \( i \) and \( \beta_{Y_i} \) is the additive change in the expectation of \( Y \) caused by a unit increase in \( X \) in individual \( i \). From this notation, the ACE is defined as \( E[\beta_Y] \) — that is, the average of \( \beta_Y \) in the population. This definition incorporates the case of a multivalued \( X \) (excluding the case where \( X \) is an unordered multivalued variable or \( X \) is a discrete variable with nonlinear effects on \( Y \), since derivatives are not defined in these cases), where the distribution of \( X \) in the population could affect the ACE. This quantity is sometimes referred to as the average derivative effect.18,19 Finally, the conventional IV estimand known as the Wald estimand (here denoted as \( \beta_{IV} \)) is defined as \( \beta_{IV} = \frac{\text{cov}(X,Z)}{\text{var}(X)} \).

The No Simultaneous Heterogeneity Assumption

We define the no simultaneous heterogeneity assumption as a combination of two conditions (Assumptions 1 and 2, defined below). The name refers to the fact that, if these two assumptions hold, then \( \beta_Y \perp (Z, \beta_X) \) — that is, heterogeneity in the causal effect is independent of heterogeneity in the instrument effect (and of the instrument). Since \( \beta_X \) and \( \beta_Y \) denote effects on the additive scale, no simultaneous heterogeneity...
focuses on additive effect modification. Of note, there is no assumption regarding multiplicative effect modification other than what is implied by our assumptions on additive effect modification (in the eAppendix; http://links.lww.com/EDE/C13 we discuss implications of nonlinear effects and nonlinear data-generating models).

Theorem 1: If the no simultaneous heterogeneity assumption holds, then $\beta_Y = \text{ACE}$ (proof in the eAppendix; http://links.lww.com/EDE/C13).

We now define the conditions for no simultaneous heterogeneity to hold using causal diagrams (see the eAppendix; http://links.lww.com/EDE/C13 for equivalent definitions using nonparametric structural equation models). For a precise articulation, it is useful to partition $U$ in Figure 1 into six nonoverlapping, exhaustive sets of variables (Table 1).

Figure 2A illustrates possible causal relationships compatible with Figure 1 (i.e., compatible with the core IV assumptions) among $U$, $Z$, $X$, $\beta_X$, and $\beta_Y$. Of note, Figure 1 assumes that $Z$ is a causal instrument, but this is not necessary for no simultaneous heterogeneity to be defined or hold. In Theorem 1, $\beta_X$ can be replaced by $\beta_Y$, here denoting the individual-level association of a noncausal instrument $Z^*$ and $X$. By noncausal instrument, we mean that $Z^*$ is not a cause of $X$. Still, it is associated with $X$ through paths that include $Z$ as a noncollider (more specifically, paths of the form $Z^* \leftarrow W \rightarrow Z \text{ or } Z^* \rightarrow [C] \leftarrow Z$, where $C$ is being conditioned on). Therefore, both modifiers of the effect of $Z$ on $X$ and modifiers of the association between $Z^*$ and $Z$ will be modifiers of the association between $Z^*$ and $X$. Modifiers of $Z$-$X$ must be independent of modifiers of $X$-$Y$ for no simultaneous heterogeneity to hold, as discussed above. Moreover, modifiers of $Z$-$Z^*$ must be independent of any cause of $Y$ (otherwise $Z^*$ would be an invalid IV). Therefore, if Assumptions 1 and 2 (see below) hold for $Z$, they also hold for $Z^*$. Given this clarification, $Z$ will be depicted as a causal instrument in Figure 2 for simplicity.

Although it is not usual to depict individual-level effects as nodes in a causal graph, $\beta_X$ and $\beta_Y$ are indeed random variables. Since $X$ is fixed in time, these two variables are simply functions of other individual-level random variables, so they are not qualitatively different from $X$ or $Y$ for example. This contrasts with variables that may also depend on non-individual level characteristics, such as person-time, which depends on follow-up duration and can be influenced by study design. Causes of $\beta_X$ can be interpreted as modifiers of the effect of $Z$ on $X$, while causes of $\beta_Y$ can be interpreted as modifiers of the effect of $X$ on $Y$. A more comprehensive description on representing individual-level effects in causal graphs is available elsewhere.

Table 2A illustrates all unmeasured variables $U_1$ to $U_6$ are d-connected to one another due to a latent variable, thus allowing for statistical dependencies between them in unrestricted ways. Although these variables could be d-connected due to other causal structures (e.g., a path $U_1 \rightarrow U_2$), this would violate the classification proposed in Table 1 (in this example, $U_1$ would be an effect modifier of both the effect of $Z$ on $X$ and the effect of $X$ on $Y$—that is, it would be a component of $U_6$), and thus blur the distinct implications of distinct types of effect modifiers. Although this is not central to our arguments, it is instructive to clarify that, from a statistical perspective, both $U_3$ and $U_4$ (for example) are effect modifiers of $\beta_X$ if $U_3$ and $U_4$ are d-connected, because $\beta_X$ will vary between strata of $U_3$ and between strata of $U_4$ (the same reasoning applies to other unmeasured variables with respect to $\beta_X$ and/or $\beta_Y$). For clarity, we use “effect modifier” to refer to variables that themselves exert effect modification and “surrogate effect modifier” to refer to variables that are d-connected with effect modifiers but do not themselves exert effect modification. Although our arguments could ignore surrogate effect modifiers, defining different types of effect modifiers allows for a more comprehensive articulation of conditions sufficient for no simultaneous heterogeneity to hold.

Since d-separation implies statistical independence, causal diagrams can be used to find conditions under which $\beta_Y$ is d-separated from (and thus statistically independent of) $\beta_X$ and $Z$. Under such conditions, no simultaneous heterogeneity holds. In Figure 2A, $\beta_Y$ is d-connected through multiple paths to both $Z$ and $\beta_X$, so $\beta_Y \perp (Z, \beta_X)$ will not generally hold.

We now describe the assumptions that define no simultaneous heterogeneity.

Assumption 1: All unmeasured variables that modify $X$-$Y$ are independent of all unmeasured variables that modify $Z$-$X$.

TABLE 1. Subsets of unmeasured variables collectively represented as $U$ in Figure 1.

| Variable | Causes $X$ | Causes $Y$ | Modifies $Z$-$X$ | Modifies $X$-$Y$ |
|----------|------------|------------|-----------------|-----------------|
| $U_1$    | Yes        | No         | Yes             | No              |
| $U_2$    | No         | Yes        | No              | Yes             |
| $U_3$    | Yes        | Yes        | No              | No              |
| $U_4$    | Yes        | Yes        | Yes             | No              |
| $U_5$    | Yes        | Yes        | No              | Yes             |
| $U_6$    | Yes        | Yes        | Yes             | Yes             |

*Strictly speaking, $U_1$ and $U_2$ do not belong to $U$ in Figure 1, but we will use the notation $U^*$ to refer to unmeasured variables in general.

*This refers to additive effect modification.
This assumption implies that there are no unmeasured variables that themselves modify (on the additive scale) both the effect of \( Z \) on \( X \) and the effect of \( X \) on \( Y \)—that is, \( U_6 = \emptyset \). Furthermore, all unmeasured variables that modify the effect of \( Z \) on \( X \) (i.e., \( U_1 \) and \( U_4 \)) are independent of all unmeasured variables that modify the effect of \( X \) on \( Y \) (i.e., \( U_2 \) and \( U_5 \)). This implies that there are neither unmeasured effect modifiers of \( Z-X \) that are also surrogate effect modifiers of \( X-Y \) nor unmeasured surrogate effect modifiers of \( Z-X \) that are also effect modifiers of \( X-Y \). However, \( U_1, U_3, \) and \( U_4 \) may be correlated with one another, and \( U_2, U_5, \) and \( U_5 \) may be correlated with one another. For this to hold, \( X \) cannot cause \( U_2 \). Otherwise, \( U_1 \) and \( U_4 \) (which are modifiers of the effect of \( Z \) on \( X \)) and \( U_2 \) (which is a modifier of the effect of \( X \) on \( Y \)) will be d-connected through the paths \( U_1 \rightarrow X \rightarrow U_2 \) and \( U_4 \rightarrow X \rightarrow U_2 \). This assumption is violated in Figure 2A, where all unmeasured variables are allowed to be d-connected with one another.

Even if Assumption 1 holds, paths of the form \( \beta_X \leftarrow Z \rightarrow X \rightarrow \beta_Y \), which render \( \beta_X \) and \( \beta_Y \) d-connected, may still exist. Therefore, Assumption 1 is necessary, but not sufficient, to render \( \beta_X \) and \( \beta_Y \) d-separated.

Assumption 2: The effect of \( X \) on the expectation of \( Y \) is additive linear.

This assumption holds if \( F_{Y_1}(x) = \varphi_i + \mu_i x \), where \( \varphi_i \) and \( \mu_i \) may vary between individuals. In this case, \( F_{Y_1}(x) - F_{Y_1}(x') = \mu_i (x - x') \). This implies that, for a given individual, the additive change in the expectation of \( Y \) caused by a unit increase in \( X \) does not depend on the value of \( X \) that the individual has. That is, the effect of \( X \) on \( Y \) is additive linear (but not necessarily constant) across individuals. This implies that the path \( X \rightarrow \beta_Y \) does not exist. Consequently, \( Z \) and \( \beta_Y \) are d-separated (that is, \( \beta_Y \perp \perp Z \)). Of note, this assumption is automatically satisfied if \( X \) is binary.

In Figure 2B, both Assumptions 1 and 2 hold. In this graph, \( \beta_Y \) and \( \beta_X \) are d-separated (because all paths from \( \beta_X \) to \( \beta_Y \) contain at least one collider), and \( \beta_Y \) and \( Z \) are d-separated (because all paths from \( Z \) to \( \beta_Y \) contain \( X \) as a collider).

Even though the above focused on the structural interpretation of no simultaneous heterogeneity (i.e., an interpretation where concepts can be represented in causal graphs), this assumption can be relaxed in the sense that it does not require full independence, but only mean independence (i.e., uncorrelatedness). That is, if \( \mathbb{E} [\beta_Y | Z, \beta_X] = \mathbb{E} [\beta_Y] \), then \( \beta_Y = \text{ACE} \) (see eAppendix; http://links.lww.com/EDE/C13 for details). Therefore, no simultaneous heterogeneity is a statistical statement, so representing it using causal graphs may be useful, but not strictly required. For example, in Figure 2A, even though no simultaneous heterogeneity will not hold in general, it may hold under lack of faithfulness (i.e., when there is d-connection but no statistical association). In this sense, no simultaneous heterogeneity is agnostic to whether faithfulness is assumed.
No Simultaneous Heterogeneity is a Generalization of Previous IV4 Assumptions

We now show that well-known assumptions that identify the ACE imply no simultaneous heterogeneity.

Causal Effect Homogeneity

For a binary $X$ and assuming deterministic counterfactuals, this assumption can be defined as $Y_i(X_i = 1) - Y_i(X_i = 0) = c$ (a constant), where $Y_i(X_i = x) = f_Y(\text{do}(X = x), U = U_i, \varepsilon_Y = \varepsilon_{Yi})$ for $x \in \{0, 1\}$. More generally, this condition can be defined as $\beta_Y = c$. Since $\beta_Y$ is constant, no simultaneous heterogeneity trivially holds—that is, it is implied by causal effect homogeneity. Moreover, no simultaneous heterogeneity allows identification when there is causal effect heterogeneity; therefore, it is weaker than causal effect homogeneity.

Instrument Effect Homogeneity

For a binary causal instrument $Z$ and assuming deterministic counterfactuals, this assumption can be defined as $X_i(Z_i = 1) - X_i(Z_i = 0) = c$, where $X_i(Z_i = z) = f_X(\text{do}(Z = z), U = U_i, \varepsilon_X = \varepsilon_{Xi})$ for $x \in \{0, 1\}$. Generally, this condition can be defined as $\beta_X = c$. Since $\beta_X$ is constant, $\beta_Y \perp \beta_X$ trivially holds. However, instrument effect homogeneity does not imply $\beta_Y \perp Z$, because the effect of $X$ on $Y$ may be nonlinear, except if $X$ is binary. This is important because additive linearity in the effect of $X$ on $Y$ is required for the conventional IV estimand to equal the ACE (see eAppendix, http://links.lww.com/EDE/C13 for details).

Therefore, instrument effect homogeneity implies that Assumption 1 is true. However, ACE identification also requires Assumption 2, which means no simultaneous heterogeneity holds. Since the latter allows identification under instrument effect heterogeneity and both no simultaneous heterogeneity and instrument effect homogeneity require Assumption 2, no simultaneous heterogeneity is weaker than instrument effect homogeneity.

No Effect Modification by Unmeasured Factors

Homogeneity can be relaxed by considering a condition sometimes referred to as no effect modification. We consider two versions of this assumption: no effect modification (1) and (2). For a binary $X$, no effect modification (1) is defined as $E[Y_i(X_i = 1) - Y_i(X_i = 0) | U_i] = E[Y_i(X_i = 1) - Y_i(X_i = 0)]$. More generally, it postulates that $E[\beta_Y | U_i] = E[\beta_Y]$—that is, no unmeasured $X$-$Y$ confounder modifies the additive effect of $X$ on the expectation of $Y$. For a binary $X$, no effect modification (1) holds if $U$ and $\beta_Y$ are d-separated; otherwise, it would not hold in general. Therefore, Assumption 1 is necessary for it to hold; otherwise, there will be open paths between $\beta_Y$ and (components of) $U$. However, Assumption 1 is not sufficient since it allows for confounders to be effect modifiers. If $X$ is continuous, Assumption 2 is also necessary (but not sufficient), otherwise, the path $X \rightarrow \beta_Y \rightarrow U$ will exist, which would render $\beta_Y$ and $U$ d-connected, thus violating no effect modification (1). Since Assumptions 1 and 2 are both necessary for no effect modification (1) and sufficient for no simultaneous heterogeneity to hold, no effect modification (1) implies no simultaneous heterogeneity. However, since these two assumptions are insufficient for no effect modification (1) to hold, no simultaneous heterogeneity does not imply no effect modification (1). Therefore, no simultaneous heterogeneity is weaker than no effect modification (1). Of note, some authors refer to violation of no effect modification (1) as essential heterogeneity.23

Although not typically referred to this way, no effect modification also applies to the association between $Z$ and $X$; we will call this condition no effect modification (2). For a binary causal instrument $Z$, this assumption is defined as $E[X_i(Z_i = 1) - X_i(Z_i = 0) | U_i] = E[X_i(Z_i = 1) - X_i(Z_i = 0)]$. More generally, no effect modification (2) postulates that $E[\beta_X | U_i] = E[\beta_X]$—that is, no unmeasured $X$-$Y$ confounder modifies the additive association between $Z$ and $X$. If both $Z$ and $X$ are binary, this condition holds if $U$ and $\beta_X$ are d-separated; otherwise, it would not hold in general. Therefore, Assumption 1 is necessary for no effect modification (2) to hold; otherwise, there will be open paths between $\beta_Y$ and (components of) $U$. However, Assumption 1 is not sufficient, since it allows for confounders to be effect modifiers. For a continuous $X$, no effect modification (2) is not sufficient to identify the ACE since even the stronger condition of instrument effect homogeneity requires Assumption 2. Since Assumptions 1 and 2 are both necessary for no effect modification (2) to identify the ACE and sufficient for no simultaneous heterogeneity to hold, no effect modification (2) implies no simultaneous heterogeneity. However, since these two assumptions are insufficient for no effect modification (2) to hold, no simultaneous heterogeneity does not imply no effect modification (2). Therefore, no simultaneous heterogeneity is weaker than no effect modification (2).

No Unmeasured Effect Modification

For binary $Z$ and $X$, it has been shown that the usual IV estimand equals the ACE if $\text{Cov}(E[Y | U, Z = 1] - E[Y | U, Z = 0], E[Y | U, \text{do}(X = 1)] - E[Y | U, \text{do}(X = 1)]) = 0$. In words, this condition (which is clearly weaker than no effect modification), postulates that there are no unmeasured variables that modify (on the additive scale) both the $Z$-$X$ association and the $X$-$Y$ effect.10 This zero-covariance condition can be expressed as $E[\beta_Y | Z, \beta_X] = E[\beta_Y]$, which is equivalent to no simultaneous heterogeneity for binary $X$ and $Z$ (this is because, for a binary $X$, $E[\beta_Y | Z, \beta_X] = E[\beta_Y | \beta_X]$ since in this case Assumption 2 is guaranteed to hold). Therefore, no simultaneous heterogeneity generalizes no unmeasured effect modification to situations involving nonbinary $Z$ and $X$. 

© 2023 Wolters Kluwer Health, Inc. All rights reserved.
SIMULATION STUDY

We performed a simulation to demonstrate further no simultaneous heterogeneity is sufficient for ACE identification (a detailed description is provided in the eAppendix, http://links.lww.com/EDE/C13). The results are shown in Figure 3. In scenario 1, no simultaneous heterogeneity holds, and the two-stage least squares estimator (equivalent to the Wald estimator for a single $Z$, $X$, and $Y$) consistently estimates the ACE, with coverage being approximately 95% and bias converging to zero as sample size increases. A similar pattern was seen when no simultaneous heterogeneity holds, but error terms were non-normal (scenarios 4 and 5). When no simultaneous heterogeneity is violated (scenarios 2 and 3), two-stage least squares had substantial bias and undercoverage. eTable 1, http://links.lww.com/EDE/C13, shows results for different two-stages least squares specifications. When no simultaneous heterogeneity is violated, but Assumption 2 holds (as in scenario 2), adjusting for measured common

![Figure 3](http://links.lww.com/EDE/C13)
effect modifiers mitigates bias. However, when no simultaneous heterogeneity is violated exclusively due to Assumption 2 being invalid (as in scenario 3), covariate adjustment does not improve estimation.

**DISCUSSION**

This paper shows that the Wald estimator is consistent for the ACE if, in addition to the core IV assumptions, the no simultaneous heterogeneity assumption holds. This condition is weaker than previously proposed IV4 assumptions that identify the ACE. This does not include the monotonicity assumption (defined precisely in the eAppendix http://links.lww.com/EDE/C13), which is sometimes classified as an IV4 assumption but does not identify the ACE. While no simultaneous heterogeneity is not strictly weaker or stronger than monotonicity, no simultaneous heterogeneity has two advantages over monotonicity. First, the latter only identifies local average causal effects, whereas no simultaneous heterogeneity identifies the ACE. Second, the no simultaneous heterogeneity assumption identifies the ACE even if there are defiers. However, if no simultaneous heterogeneity is violated but monotonicity holds, IV estimators, identify the local average causal effect, which is a well-defined causal parameter for binary treatments. Moreover, the notion of compliers is not well-defined for continuous treatments. In this case, monotonicity allows interpreting the Wald estimator as a weighted average of treatment effects, with subgroups of the population where the \( Z \times X \) association is stronger receiving greater weight.20,26 Although mathematically well-defined, this parameter is difficult to interpret for policy making. Conversely, if no simultaneous heterogeneity holds, then IV estimators will identify the ACE of a continuous treatment.

In many recent papers describing methodological developments that relax the exclusion restriction assumption when there are multiple IVs, treatment effect homogeneity is implicitly or explicitly required.27,28 This is because these methods assume that valid IVs identify the same causal effect (generally, the ACE). However, assuming no simultaneous heterogeneity holds for all IVs is sufficient to identify the ACE. This implies that the assumptions required for the validity of these methods are weaker than previously considered. Nevertheless, since we defined no simultaneous heterogeneity for a time-fixed treatment, caution must be taken to extrapolate our conclusions to time-varying treatments. Moreover, even though no simultaneous heterogeneity is weaker than homogeneity or no effect modification, it is still quite strong and should not be taken for granted in practice.

No simultaneous heterogeneity is an untestable assumption that cannot be guaranteed by study design. Therefore, assessing its plausibility requires subject-matter knowledge. We illustrate this by discussing three published IV studies (eAppendix; http://links.lww.com/EDE/C13). The possibility of (partially) empirically verifying some IV4 assumptions has implications for no simultaneous heterogeneity plausibility. For example, in the case of a continuous treatment, instrument effect heterogeneity would often imply (except in some specific circumstances) that the treatment is heteroscedastic with respect to the instrument (i.e., the variance of the treatment would differ between levels of the instrument).29,30 Future methodologic studies are required to assess the power and utility of such tests in typical IV settings. Since no simultaneous heterogeneity is weaker than homogeneity, homoscedasticity, in this case, would support (but not guarantee) that no simultaneous heterogeneity holds. However, heteroscedasticity would not necessarily imply that this assumption is violated. Brookhart (2007)8 proposed empirically assessing the plausibility of instrument effect homogeneity by testing if instrument strength varies between strata of measured covariates. This can also be viewed as a test of no simultaneous heterogeneity since if multiple covariates modify instrument strength, then the assumption that there are no unmeasured common effect modifiers is less plausible. Indeed, assuming \( Z \) is a valid instrument, variability in IV estimates between strata defined by such covariates could be interpreted as evidence against no simultaneous heterogeneity. Strategies to empirically verify this assumption remain to be formally investigated.

An earlier glimpse of no simultaneous heterogeneity in the literature can be found, for example, in Angrist,31 who noted that variation in response to the draft (i.e., \( Z \times X \) heterogeneity) based on potential outcomes (i.e., \( X \times Y \) heterogeneity) could mean his results were biased estimates of the ACE. Indeed, this would be a scenario where no simultaneous heterogeneity is violated. For binary \( Z \) and \( X \), no simultaneous heterogeneity is equivalent to no unmeasured effect modification.10 However, no causal structures (e.g., in the form of causal graphs) that dictate whether no simultaneous heterogeneity is violated were presented. Syrgkanis et al.13 also postulated an independence condition that is equivalent to no simultaneous heterogeneity, without necessarily restricting to binary \( Z \) and \( X \). However, mechanisms that could render this condition satisfied were similarly not discussed, and no explicit consideration was given to potential implications of non-linear instrument-outcome associations or treatment-outcome effects. Moreover, their assumed data-generating mechanism for the outcome was \( Y_i = \theta(U_i)X_i + \rho(U_i) + \epsilon_i \), where \( \theta(U) \) is the causal effect function, which was assumed to be linear additive. However, such a model may not be appropriate when \( Y \) is binary, which is often assumed to be a non-linear function of \( X \) and \( U \) (e.g., the expectation of \( Y \) may have a logistic function of \( X \) and \( U \)).

Here we use counterfactual notation to explicitly define individual-level effects in a framework that allows for both binary and continuous \( Z \), \( X \), and/or \( Y \). We also comprehensively describe data-generating mechanisms influencing no simultaneous heterogeneity using causal graphs. This helps to apply expert knowledge to assess the plausibility of this assumption in practice. We also explicitly consider the implications of nonlinear \( Z \times X \) associations and \( X \times Y \) effects.
The present work thus clarifies the assumptions underlying no simultaneous heterogeneity, which allows differentiating it from previous IV4 assumptions and explicitly propose no simultaneous heterogeneity as an IV4 assumption that is weaker than previously described IV4 assumptions that identify the ACE.

REFERENCES

1. Angrist JD, Krueger AB. Instrumental variables and the search for identification: from supply and demand to natural experiments. J Econ Perspect. 2001;15:69–85.
2. Bollen KA. Instrumental variables in sociology and the social sciences. Annu Rev Sociol. 2012;38:37–72.
3. Ertefaie A, Small DS, Flory JH, Hennessy S. A tutorial on the use of instrumental variables in pharmacoepidemiology. Pharmacoepidemiol Drug Saf. 2017;26:357–367.
4. Hernán MA, Robins JM. Instruments for causal inference: an epidemiologist's dream? Epidemiology. 2006;17:360–372.
5. Didelez V, Sheehan N. Mendelian randomization as an instrumental variable approach to causal inference. Stat Methods Med Res. 2007;16:309–330.
6. Hernán MA, Robins JM. Causal Inference: What If? Chapman & Hall/ CRC; 2020.
7. Heckman J. Instrumental variables: a study of implicit behavioral assumptions used in making program evaluations. J Hum Resour. 1997;32:441–462.
8. Brookhart MA, Schneeweiss S. Preference-based instrumental variable methods for the estimation of treatment effects: assessing validity and interpreting results. Int J Biostat. 2007;3:Article 14.
9. Wang L, Tchetgen Tchetgen E. Bounded, efficient and multiply robust estimation of average treatment effects using instrumental variables. J R Stat Soc Series B Stat Methodol. 2018;80:531–550.
10. Cui Y, Tchetgen Tchetgen E. A semiparametric instrumental variable approach to optimal treatment regimes under endogeneity. J Amer Statist Assoc. 2020;116:162–173.
11. Angrist JD, Imbens GW, Rubin DB. Identification of causal effects using instrumental variables. J Am Stat Assoc. 1996;91:444–455.
12. Harris KM, Remler DK. Who is the marginal patient? Understanding instrumental variables estimates of treatment effects. Health Serv Res. 1998;33:1337–1360.
13. Syrgkanis V, Lei V, Oprescu M, Hei M, Battocchi K, Lewis G. Machine Learning Estimation of Heterogeneous Treatment Effects with Instruments. NIPS’19: Proceedings of the 33rd International Conference on Neural Information Processing Systems. 2019;15193–15202. Available at: https://dl.acm.org/doi/pdf/10.5555/3454287.3455648.
14. Imbens GW. Better LATE than nothing: some comments on Deaton (2009) and Heckman and Urzua (2009). J Econ Lit. 2010;48:399–423.
15. Aronow PM, Carnegie A. Beyond LATE: estimation of the average treatment effect with an instrumental variable. Political Anal. 2013;21:492–506.
16. Angrist JD, Fernández-Val I. ExtrapoLATE-ing: external validity and overidentification in the LATE framework. In: Acemoglu D, Arellano M, Dekel E. Advances in Economics and Econometrics: Tenth World Congress. Cambridge (England): Cambridge University Press; 2011. 401–434.
17. Vansteelandt S, Didelez V. Improving the robustness and efficiency of covariate-adjusted linear instrumental variable estimators. Scand J Stat. 2018;45:941–961.
18. Hardle W, Stoker TM. Investigating smooth multiple regression by the method of average derivatives. J Am Stat Assoc. 1989;84:986–996.
19. Powell JL, Stock JH, Stoker TM. Semiparametric estimation of index coefficients. J Econ. 1989;57:14031430.
20. Angrist JD, Imbens GW. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. J Am Stat Assoc. 1995;90:431–442.
21. Nilsson A, Bonander C, Strömberg U, Björk J. A directed acyclic graph for interactions. Int J Epidemiol. 2021;50:613–619.
22. Hartwig FP, Wang L, Smith GD, Davies NM. Homogeneity in the instrumental-treatment association is not sufficient for the Wald estimator to equal the average causal effect when the exposure is continuous. Epidemiology. 2022;33:828–831.
23. Basu A, Heckman JJ, Navarro-Lozano S, Urzua S. Use of instrumental variables in the presence of heterogeneity and self-selection: an application to treatments of breast cancer patients. Health Econ. 2007;16:1133–1157.
24. Angrist JD, Graddy K, Imbens GW. The interpretation of instrumental variables estimators in simultaneous equations models with an application to the demand for fish. Rev Econ Stud. 2000;67:499–527.
25. Angrist JD, Pischke J-S. Chapter 4. Instrumental Variables in Action: Sometimes You Get What You Need. In: Mostly Harmless Econometrics. Princeton University Press; 2009. 113–220.
26. von Hinke S, Davey Smith G, Lawlor DA, Propper C, Windmeijer F. Genetic markers as instrumental variables. Int J Epidemiol. 2016;40:304–314.
27. Bowden J, Davey Smith G, Haycock PC, Burgess S. Consistent estimation in mendelian randomization with some invalid instruments using a weighted median estimator. Genet Epidemiol. 2016;45:131–148.
28. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. J Epidemiol. 2017;46:1985–1998.
29. Mills HL, Higgins JPT, Morris RW, et al. Detecting heterogeneity of intervention effects using analysis and meta-analysis of differences in intervention effects. J Health Serv Res. 2020;11:45:304–314.
30. Hartwig FP, Smith GD, Bowden J. Robust inference in summary data mendelian randomization via the zero modal pleiotropy assumption. J Epidemiol. 2017;46:1985–1998.
31. Angrist JD. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. Amer Econ Rev. 1990;80:313–336.