Depicting deterministic variables within directed acyclic graphs (DAGs): An aid for identifying and interpreting causal effects involving derived variables and compositional data.

*Laurie Berrie[1], *Kellyn F Arnold[2], Georgia D Tomova[2,3,4], Mark S Gilthorpe[4,5], Peter WG Tennant[2,3,4]

[1]School of GeoSciences, University of Edinburgh, Edinburgh, UK; [2]Leeds Institute for Data Analytics, University of Leeds, Leeds, UK; [3]School of Medicine, University of Leeds, Leeds, UK; [4]The Alan Turing Institute, London, UK; [5]Obesity Institute, Leeds Beckett University, Leeds, UK.

*Lead authors

Corresponding author

Peter WG Tennant PhD

Address: Leeds Institute for Data Analytics,
Level 11 Worsley Building,
Clarendon Way,
Leeds,
LS2 9NL,
UK.

Email: P.W.G.Tennant@leeds.ac.uk

Twitter: @PWGTennant

Acknowledgments: We thank Sarah Gadd for her contributions to some of the ideas presented in this paper, Sarah Wieten for reading and commenting on the manuscript, and the anonymous peer reviewers for their helpful suggestions.

Funding: This work received no specific funding, but during the conduct of this work LB received funding from the Medical Research Council [MR/K501402/1], KFA received funding from the Economic and Social Research Council [ES/J500215/1] and KFA, GDT, MSG, and PWGT received funding from the Alan Turing Institute [EP/N510129/1].

Conflict of Interest: PWGT is a director of Causal Thinking Ltd and MSG is a director of Causal Insights Solutions Ltd, both of which provide causal inference research and training. Both companies and their directors may benefit from any study that demonstrates the value of causal inference methods.

ABSTRACT

Deterministic variables are variables that are functionally determined by one or more parent variables. They commonly arise when a variable has been functionally created from one or more parent variables, as with derived variables, and in compositional data, where the 'whole' variable is determined from its 'parts'.

This article introduces how deterministic variables may be depicted within directed acyclic graphs (DAGs) to help with identifying and interpreting causal effects involving derived variables and/or compositional data. We propose a two-step approach in which all variables are initially considered, and a choice is made whether to focus on the deterministic variable or its determining parents.

Depicting deterministic variables within DAGs brings several benefits. It is easier to identify and avoid misinterpreting tautological associations, i.e., self-fulfilling associations between deterministic variables and their parents, or between sibling variables with shared parents. In compositional data, it is easier to understand the consequences of conditioning on the ‘whole’ variable, and correctly identify total and relative causal effects. For derived variables, it encourages greater consideration of the target estimand and greater scrutiny of the consistency and exchangeability assumptions.

DAGs with deterministic variables are a useful aid for planning and interpreting analyses involving derived variables and/or compositional data.

KEY WORDS:

Causal inference, directed acyclic graphs, compositional data, derived variables, composite variables, tautological associations

INRODUCTION

Causal directed acyclic graphs (DAGs) are increasingly popular aids for identifying and estimating causal effects(1,2) and for recognizing and understanding various forms of error, bias, and non-causal associations.(3–7) However, little attention has been given to their utility for understanding analyses involving **deterministic variables**.(8,9) A deterministic variable is a variable that is functionally determined by one or more other variables such that its value can be known with certainty once its parents are known.(8,9) They are extremely common in health and social science, typically arising in the following types and settings:

1. **Derived variables.** Derived variables are variables that have been functionally created from one or more parent variables.(10) They include **simple derived variables** (e.g., macrosomia), which are created from a single parent variable (e.g., birthweight), and **composite derived variables** (e.g., waist-to-hip ratio), which are created from two or more parent variables (e.g., waist circumference and hip circumference).(11)
2. **Compositional data**. Compositional data is a form of hierarchical data that contains ‘part variables’ (e.g., fat mass and fat-free mass) that perfectly sum to a ‘whole’ variable (e.g., total mass) or a constant.(12–14)

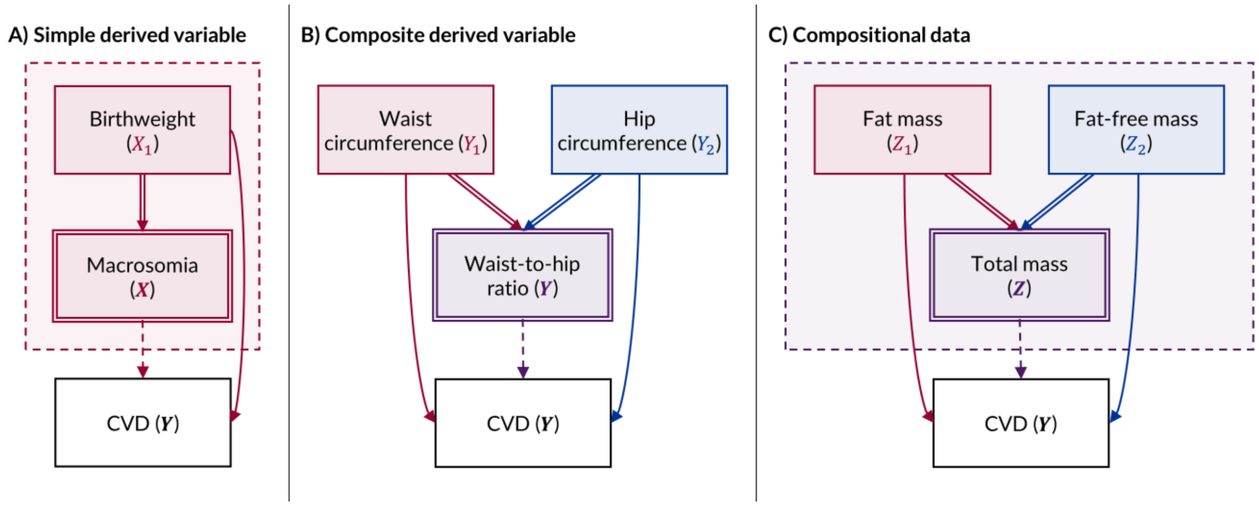
Because DAGs are primarily used to consider probabilistic relationships,(15) deterministic variables have received limited attention within DAGs. Indeed, DAGs containing deterministic variables have additional statistical implications that make them incompatible with many routine causal identification and discovery algorithms.(16) Depicting deterministic variables within DAGs can, however, be useful for understanding certain challenges involved in the analyses and interpretation of deterministic variables.

In the following, we introduce how deterministic variables can be depicted within DAGs and discuss the benefits for identifying and interpreting causal effects involving derived variables and/or compositional data.

DEPICTING DETERMINISTIC VARIABLES WITHIN DAGS

A causal DAG is a graphical representation of the hypothesized causal relationships between a set of variables (or 'nodes').(1,2) Any two variables in the graph may be connected by a unidirectional arrow (or 'arc'), which signifies that the first variable (the 'parent' or 'ancestor') exerts a causal effect on the second (the 'child' or 'descendent'). Because a DAG is acyclic, no variable may cause itself at the same moment in time. To ensure that deterministic variables are handled appropriately within DAGs, they should be distinctively depicted. To achieve this, we follow the convention that any 'child' variable that is fully determined by one or more 'parent' variables is depicted with a double-outlined node.(17) We also suggest: 1) all arcs entering a deterministic variable should be double-lined, to denote that they are part of a functional, not probabilistic, relationship(12); and 2) all arcs leaving a deterministic variable should be dashed; this denotes that, while it may be useful to conceptualize the 'implied' causal effect of a deterministic variable, no residual effect exists beyond that caused by the parent variables.(10,18) Finally, where a child variable and all determining parent variables occur concurrently, we suggest enclosing the family within a dashed-outline box.(12) Examples of this notation are given in **Figure 1**, which depicts a simple derived variable (**Figure 1A**), a composite derived variable (**Figure 1B**), and compositional data (**Figure 1C**).

Figure 1 Causal directed acyclic graphs using deterministic notation to depict: A) a simple derived variable; B) a composite derived variable; and C) compositional data.


A) Simple derived variable
Birthweight (X_1)
Macrosomia ($\boldsymbol{X}$)
CVD ($\boldsymbol{Y}$)
B) Composite derived variable
Waist circumference (Y_1)
Hip circumference (Y_2)
Waist-to-hip ratio ($\boldsymbol{Y}$)
CVD ($\boldsymbol{Y}$)
C) Compositional data
Fat mass (Z_1)
Fat-free mass (Z_2)
Total mass ($\boldsymbol{Z}$)
CVD ($\boldsymbol{Y}$)

Fully determined child variables are represented by double-outlined nodes, deterministic relationships are represented by double-lined arcs, arcs leaving a deterministic variable are represented as dashed arcs, and situations where a child variable and its determining parent variables occur simultaneously in time are enclosed within a dashed-outline box. An example outcome, cardiovascular disease (CVD) has been added to all examples. In A) the simple derived variable, macrosomia ($\boldsymbol{X}$), is a binary variable that is fully determined by birthweight (X_1). In B) the composite derived variable, waist-to-hip ratio ($\boldsymbol{Y}$), is fully determined by dividing the waist circumference (Y_1) by the hip circumference (Y_2). In C) the 'whole' variable, total mass ($\boldsymbol{Z}$), comprises two 'part' variables, fat mass (Z_1) and fat-free mass (Z_2) and can therefore be fully determined by summing both parent components.

ALGORITHMIC APPROACHES AND THE BENEFIT OF DAGS

For many years, deterministic variables were not strictly compatible with DAGs because deterministic variables bring additional statistical dependencies.(1) This was resolved with the introduction of the ***D*-separation** criterion (note the uppercase '*D*'), which extends the familiar *d*-separation criterion to accommodate the behavior of deterministic variables.(17) Despite this, most causal modelling and discovery algorithms are not natively compatible with deterministic variables.(16) Such variables are hence usually treated as nuisance nodes that need to be identified and removed.(17) Shachter's **Deterministic Node Reduction** algorithm achieves this by identifying all deterministic variables within a DAG and transferring the incoming and outgoing probabilistic arcs to their parent nodes to create **barren nodes** that may be removed from the graph without losing information about the relationships between the remaining variables.(8)

Simply identifying and removing deterministic nodes is not especially useful when a deterministic variable is the exposure or outcome of interest. In these circumstances, we advocate an alternative two-step approach to ensure that the assumptions and implications are fully considered. First, a 'full' DAG is drawn that includes the deterministic exposure and/or outcome, as well as all determining parents. Next, an explicit choice is made whether to focus on the deterministic variable(s) or the determining parents.

UNDERSTANDING TAUTOLOGICAL ASSOCIATIONS

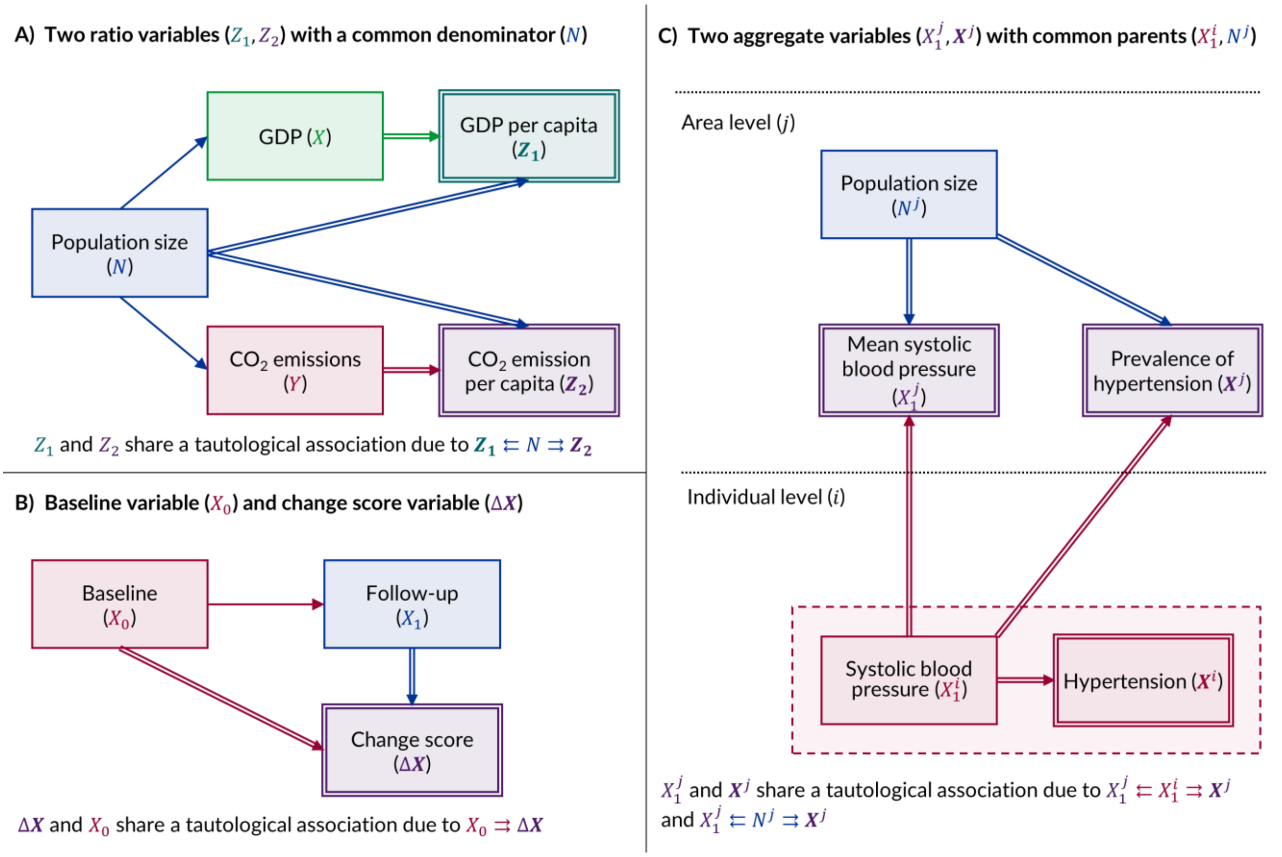
Perhaps the most straightforward benefit of depicting deterministic variables within DAGs is the ability to identify and avoid misinterpreting tautological associations. We define tautological associations as the self-fulfilling associations that arise when a deterministic variable is analyzed in direct relation to one of its parent variables, or a sibling variable with at least one shared parent component.

The problem of tautological associations was first identified by Karl Pearson in 1897 in the context of analyzing **ratio variables.**(19) Ratio variables (e.g., X/N) are composite derived variables created by dividing one parent variable (e.g., X) by a second parent variable (e.g., N).(20) Assuming faithfulness,(1) Pearson warned that two ratio variables with a shared denominator parent variable (e.g., X/N and Y/N) would share a '*spurious (organic) correlation*' even if the numerators (e.g., X, Y) are unrelated.(19) Using deterministic notation, this phenomenon can be depicted and understood using a DAG that contains the three parent variables (i.e., X, Y, and N) and the two child variables (X/N and Y/N) (**Figure 2A**).

Perhaps the most well-known example of a tautological association occurs in the context of analyzing **change score variables**. Change score variables (e.g., $\Delta X = X_1 - X_0$) are composite derived variables created by subtracting an earlier measure of a time-varying variable (e.g., X_0) from a subsequent measure of that variable (e.g., X_1) (**Figure 2B**). In 1962, Oldham warned that change score variables share a negative '*spurious correlation*' with their baseline parent variable that is '*entirely produced by our arithmetical procedure*' (21) Known sometimes as the 'law of initial value', this occurs because of the negative parametrization of the baseline variable in the change score variable.(22) Other examples of tautological associations can be found in the literature, under the term '*mathematic(al) coupling*',(23) although most examples probably occur in applied analyses with no awareness of the phenomenon.

In statistical terms, tautological associations are neither erroneous nor biased.(24) The expected association between two ratio variables, for example, is an accurate reflection of their common denominator variable.(24) However, for causal interpretation, inferential bias can occur when the underlying tautology is not recognized and the resulting associations are misattributed to other (causal) mechanisms. Such misinterpretations are probably more common for composite derived variables with many parent variables since the deterministic origins become easier to overlook. Nevertheless, there are examples of simple tautological associations (e.g., between hypertension and blood pressure) being overlooked when analyses are conducted at aggregate level (**Figure 2C**).(25) By placing the parent variables for all deterministic exposure and outcome variables within a DAG, we believe such mistakes become less likely.

Figure 2 Causal directed acyclic graphs of three tautological associations: A) between two ratio variables with a common denominator parent variable; B) between a change-score variable and its baseline parent variable; and C) between two aggregate variables.



A) Two ratio variables (Z_1, Z_2) with a common denominator (N)
GDP (X)
GDP per capita ($\boldsymbol{Z_1}$)
Population size (N)
CO_2 emissions (Y)
CO_2 emission per capita ($\boldsymbol{Z_2}$)
Z_1 and Z_2 share a tautological association due to $\boldsymbol{Z_1} \leftleftarrows N \rightrightarrows \boldsymbol{Z_2}$
B) Baseline variable (X_0) and change score variable (ΔX)
Baseline (X_0)
Follow-up (X_1)
Change score ($\boldsymbol{\Delta X}$)
$\boldsymbol{\Delta X}$ and X_0 share a tautological association due to $X_0 \rightrightarrows \boldsymbol{\Delta X}$
C) Two aggregate variables (X_1^j, $\boldsymbol{X}^j$) with common parents (X_1^i, N^j)
Area level (j)
Population size (N^j)
Mean systolic blood pressure (X_1^j)
Prevalence of hypertension ($\boldsymbol{X}^j$)
Individual level (i)
Systolic blood pressure (X_1^i)
Hypertension ($\boldsymbol{X}^i$)
X_1^j and $\boldsymbol{X}^j$ share a tautological association due to $X_1^j \leftleftarrows X_1^i \rightrightarrows \boldsymbol{X}^j$ and $X_1^j \leftleftarrows N^j \rightrightarrows \boldsymbol{X}^j$

In A) the observed variables GDP ($\boldsymbol{X}$) and CO2 emissions ($\boldsymbol{Y}$) are both caused by population size (N), making N a confounder for the apparent relationship between X and Y. Two composite derived variables have also been created by diving both X and Y by N to create GDP per capita ($\boldsymbol{Z_1}$) and CO2 emission per capita ($\boldsymbol{Z_2}$), respectively. Since both $\boldsymbol{Z_1}$ and $\boldsymbol{Z_2}$ share the same parent variable (N), they will share a tautological association. In B) the baseline measurement of a repeated measure variable (X_0) causes the follow-up measurement (X_1), from which a change score variable ($\boldsymbol{\Delta X}$) has been created by subtracting X_0 from X_1. Since X_0 is a deterministic parent of $\boldsymbol{\Delta X}$, they share a tautological association. In C), where subscript i denotes individual-level and j denotes area-level, a simple derived variable, hypertension ($\boldsymbol{X^i}$), is a dichotomized individual-level variable that is fully determined by the continuous individual-level variable systolic blood pressure ($\boldsymbol{X_1^i}$). Area-level mean systolic blood pressure ($\boldsymbol{X_1^j}$) and area-level prevalence of hypertension ($\boldsymbol{X^j}$) are determined at the aggregate level from $\boldsymbol{X_1^i}$ and the area-level population ($\boldsymbol{N^j}$). Since both $\boldsymbol{X_1^j}$ and $\boldsymbol{X^j}$ share the same two parent variables ($\boldsymbol{X_1^i}$, $\boldsymbol{N^j}$), they share a tautological association.

CONSIDERING SIMPLE DERIVED VARIABLES

Most simple derived variables are created for statistical rather than causal reasons. For example, an exposure variable may be log transformed to more accurately model a nonlinear relationship with the outcome.(26) Although important for estimation,(27) such transformations have limited implications for causal reasoning, except where the transformation leads to a loss of information, e.g., due to **coarsening**.(28)

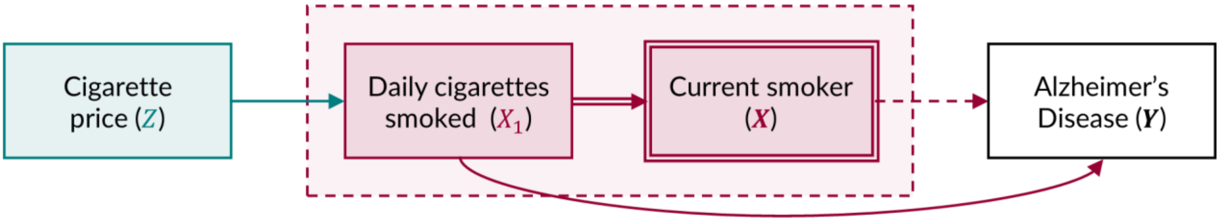
Coarsening commonly occurs when a **dichotomous variable** (e.g., smoker) is created by collapsing a continuous or multinomial variable (e.g., daily cigarettes smoked) into two categories.(29) Coarsening the exposure has particular implications for the consistency assumption, which requires that there must be no two versions of the exposure such that, for the same exposure value, the different versions have different probabilities of one or more possible outcomes.(30) If multiple versions of the exposure are collapsed into a single value, then these versions need to have the same effect on the outcome to provide a well-defined causal effect, an assumption known as **effect equivalence**.(28) Without effect equivalence, the estimated effect of the coarsened variable will be a poorly defined weighted average of the effects of the multiple parent versions of the exposure and their frequencies in the study population.(31)

Coarsening the exposure can be especially problematic for instrumental variable analyses because a coarsened exposure will likely violate the exclusion restriction assumption, i.e., that the instrument has no effect on the outcome other than through the exposure.(32,33) This is because the instrument is likely to cause the outcome through variation in the parent exposure variable that is not captured by the coarsened child exposure.(32,33)

To illustrate, suppose we are interested in the average causal effect of smoking on the risk of Alzheimer's disease. Smoking is a dichotomized child of daily cigarettes smoking. If the number of daily cigarettes smoked has a dose-response relationship with the risk of Alzheimer's disease, then the average causal effect of smoking will reflect a poorly defined weighted average of different smoking levels. If we tried to estimate this effect using an instrumental variable (e.g., cigarette price), the exclusion restriction assumption would be violated by any such dose-response effect.(32,33) This assumption can be seen visually in a DAG containing both the parent and child exposure variables as the (residual) path between the parent exposure and the outcome (**Figure 3**).

Although the issues with variable coarsening cannot be solved by simply depicting derived variables and their parent variables within DAGs, the practice may help to ensure that the resulting implications and assumptions are more explicitly considered.

Figure 3 Causal directed acyclic graph of an instrumental variable scenario with a coarsened exposure variable.


Cigarette price (Z)
Daily cigarettes smoked (X_1)
Current smoker ($\boldsymbol{X}$)
Alzheimer's Disease ($\boldsymbol{Y}$)

In this scenario, the continuous variable, daily cigarettes smoked (X_1) has been coarsened into a dichotomised variable, current smoker ($\boldsymbol{X}$). The implied causal effect of $\boldsymbol{X}$ on the outcome, Alzheimer's Disease ($\boldsymbol{Y}$) is depicted with a dashed arc, although this effect is technically entirely explained by X_1. Another direct path is depicted from the parent exposure to the outcome ($X_1 \rightarrow Y$), which represents the residual (dose-response) effect of the X_1 on the outcome that does not act through $\boldsymbol{X}$. For the causal effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ to be robustly estimated, this residual path $X_1 \rightarrow Y$ must be zero (the effect equivalence assumption). This is apparent if we imagine estimating the effect of $\boldsymbol{X}$ on $\boldsymbol{Y}$ using an instrumental variable, cigarette price ($\boldsymbol{Z}$), since the residual path $X_1 \rightarrow Y$, if non-zero, would violate the exclusion restriction assumption.

CONSIDERING COMPOSITIONAL DATA

The benefits of depicting deterministic variables within DAGs increase with the complexity of the variables and/or relationships being considered. Compositional data is a common form of complex data structure that naturally contains deterministic relationships because the 'part' variables sum to a 'whole' variable or constant (**Figure 1**).(12,14) This makes compositional data notoriously challenging to analyze and interpret correctly.(12,14,34,35) Pearson's warning on the use of ratio variables was allegedly motivated by observing biologists dividing bone measurement variables (e.g., femur length) by length measurement variables (e.g., leg length).(19) Since then, the area of **Compositional Data Analysis (CoDA)** has emerged to develop specific analytical strategies.(13,14) We focus on the insights that arise from considering compositional data within DAGs.(12,34,35)

Simplifying features of compositional data

There are two important features of compositional data that reduce the potential complications when compared with composite derived variables. First, the 'whole' variable can usually be directly observed. Indeed, whether a variable is a 'whole' or 'part' is often a matter of perspective or convenience rather than external structure. All variables can potentially be divided into further parts or summed to a greater whole. The choice of whether to focus on the 'whole' or the 'parts' is therefore usually a trade-off between the competing benefits of aggregation and subdivision.

The second key feature of compositional data is that the 'whole' variable and all 'part' variables occur at the same moment in time. This avoids many of the more serious issues affecting composite derived variables discussed below.

Choosing the target estimand

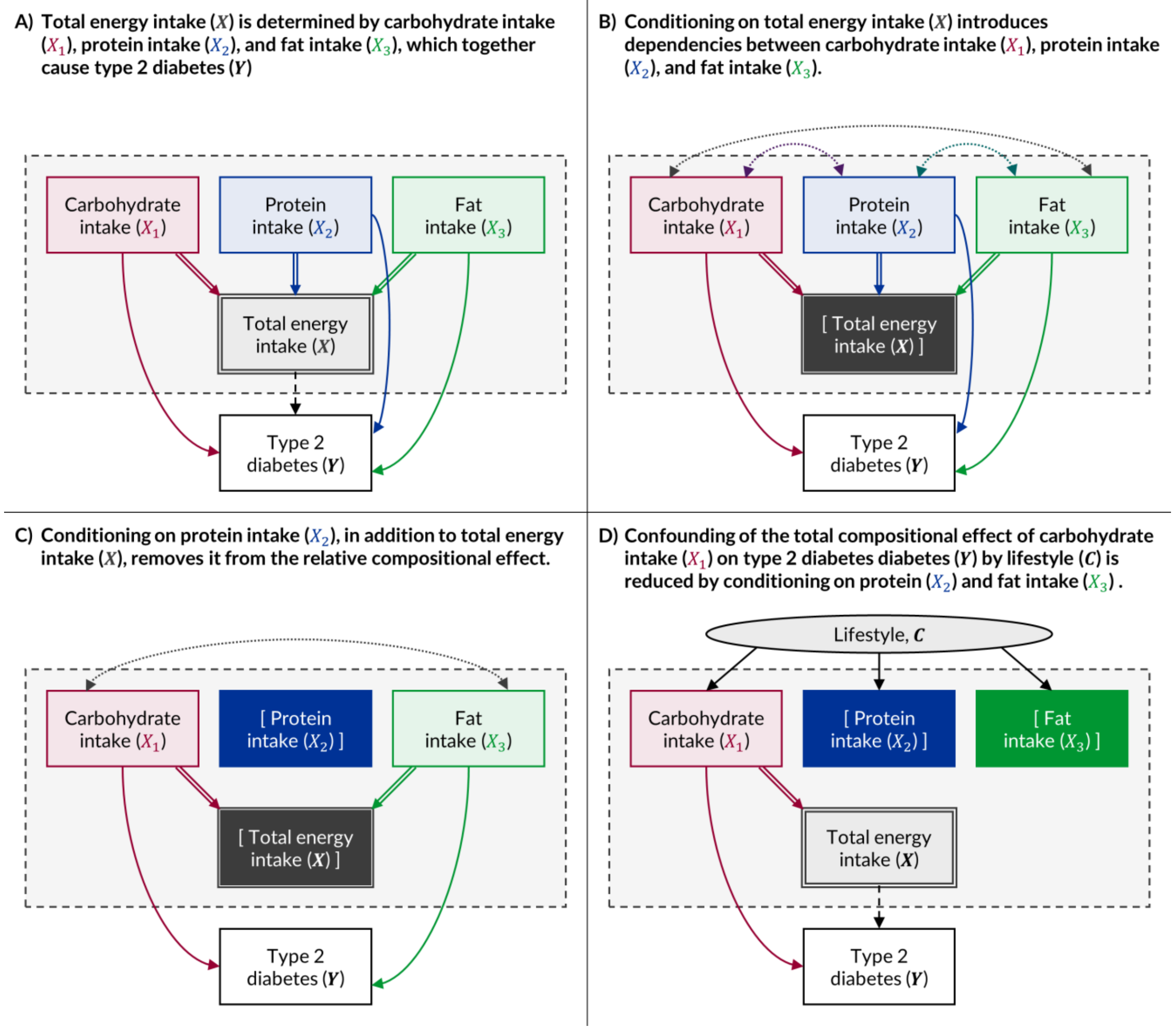
Analyses of compositional data generally consider two types of estimands: **total compositional effects** and **relative compositional effects**. Total compositional effects represent the effects of increasing the 'whole' variable either by intervening on the 'whole' directly or through one or more specified 'part' exposures.(12,34–36) Relative compositional effects represent the joint effect of increasing a specified 'part' exposure while simultaneously decreasing one or more substituting 'parts' to keep the 'whole' fixed.(12,34–36) Different analytical strategies are required to estimate these two effects, and misinterpretations occur when the wrong strategy is used inadvertently, or when the relative nature of structurally fixed data (e.g., time-use data) are not recognized.(12,34,35)

Since compositional data occur at the same time, the 'whole' and 'part' variables may be drawn in multiple ways, but it is intuitive to consider the 'whole' as being determined by the 'parts' (Figure 1C).(12) Drawn like this, the 'whole' can be usefully interpreted as a *collider* for the 'parts', and it is clear that conditioning on the 'whole' introduces a dependency between the 'parts'.(12) The individual effect of a specific 'part' cannot therefore be estimated when conditioning on the 'whole'.(12)

To illustrate, we consider the total effect of carbohydrate consumption on the risk of diabetes, where the consumption of carbohydrates, proteins, and fats determines the total energy intake (**Figure 4A).**

In nutritional epidemiology, it is common to evaluate the effects of one or more specific dietary component(s) on subsequent health outcomes while conditioning on total energy intake as a proxy "confounder" for the diet.(37) However, when drawn as suggested, it is apparent that conditioning on total energy intake would introduce a dependency between carbohydrate consumption and the other macronutrient variables, creating a relative effect (**Figure 4B**). Exactly which relative effect will depend on whether additional adjustments are made for any of the other macronutrients (**Figure 4C**).(34,35)

Figure 4 Causal directed acyclic graphs examining the identification of causal effects in compositional data, here represented by total energy intake.



A) Total energy intake (X) is determined by carbohydrate intake (X_1), protein intake (X_2), and fat intake (X_3), which together cause type 2 diabetes (Y)

B) Conditioning on total energy intake (X) introduces dependencies between carbohydrate intake (X_1), protein intake (X_2), and fat intake (X_3).

C) Conditioning on protein intake (X_2), in addition to total energy intake (X), removes it from the relative compositional effect.

D) Confounding of the total compositional effect of carbohydrate intake (X_1) on type 2 diabetes diabetes (Y) by lifestyle (C) is reduced by conditioning on protein (X_2) and fat intake (X_3) .

Carbohydrate intake (X_1)
Protein intake (X_2)
Fat intake (X_3)
Total energy intake (X)
Type 2 diabetes ($\boldsymbol{Y}$)
Carbohydrate intake (X_1)
Protein intake (X_2)
Fat intake (X_3)
[Total energy intake ($\boldsymbol{X}$)]
Type 2 diabetes ($\boldsymbol{Y}$)
Carbohydrate intake (X_1)
[Protein intake (X_2)]
Fat intake (X_3)
[Total energy intake ($\boldsymbol{X}$)]
Type 2 diabetes ($\boldsymbol{Y}$)
Lifestyle, $\boldsymbol{C}$
Carbohydrate intake (X_1)
[Protein intake (X_2)]
[Fat intake (X_3)]
Total energy intake ($\boldsymbol{X}$)
Type 2 diabetes ($\boldsymbol{Y}$)

The 'full' DAG for the scenario is depicted in A), where the 'whole' variable, total energy intake ($\boldsymbol{X}$) is fully determined by intake from three 'part' variables, carbohydrate intake (X_1), protein intake (X_2), and fat intake (X_3), which together cause type 2 diabetes ($\boldsymbol{Y}$). In B) the 'whole' has been conditioned on (denoted by square brackets and an absence of onward arcs), inducing conditional dependencies between the unconditioned 'part' variables. The causal effect of any 'part' variable (e.g., X_1) on $\boldsymbol{Y}$ would thus be relative to the other unconditioned 'part' variables (e.g., X_2, and X_3). In C) one of the other 'part' variables (X_2) has been conditioned on in addition to the whole, which removes it from the relative compositional effect; the effect of X_1 on $\boldsymbol{Y}$ will thus be relative to X_3 only. In D) a confounding variable, lifestyle ($\boldsymbol{C}$), is introduced that commonly causes all 'part' variables. Confounding from such common causes can be reduced by conditioning on other 'part' variables to block the confounding paths downstream.

Identifying and estimating causal effects in compositional data

This simple three-nutrient example demonstrates how retaining both the 'whole' and 'part' variables within a DAG can help with understanding compositional data. The optimal analytical strategy then depends on whether the investigator is interested in the total effect of a particular 'part', a relative effect of a particular 'part', or the summary effect of the 'whole'. There are, however, additional caveats.

First, the exchangeability assumption requires that the units of analysis have an equal probability of all possible values of the outcome at the time of exposure, i.e., there must be no confounding or selection bias for the exposure-outcome relationship of interest.(1,3) In compositional data, confounding can arise from common causes of the ‘parts’, even if these causes have no residual effect on the outcome, because each part is itself likely to cause the outcome. In our three-nutrient example, such common causes might include lifestyle behaviors. Ideally, these common causes should be directly measured and conditioned on, but this is not always possible. Instead, confounding by common causes may be reduced by conditioning on other 'parts' to block the confounding paths downstream (**Figure 4D**). Where each 'part' has a unique effect and variance, this requires measuring and conditioning on every part variable. In practice, aggregated 'part' variables are often used, such as 'remaining energy intake' (i.e., energy from all parts *except* the exposure), but this may introduce residual confounding wherever the causal effect of each 'part' differs from the average effect of the aggregate variable.(34,35)

In some situations, the average effect of increasing the 'whole' may be of more interest than the individual 'parts' specifically. Here, it may be reasonable to discard the parent variables from the DAG and treat the 'whole' variable as the exposure. However, this increases the chance of violating the consistency assumption, since the same value of the 'whole' can be obtained from many different combinations of the 'parts'. If each 'part' has different causal effects on the outcome and/or different variances, then the summary effect of the 'whole' will not be the mean-weighted average effect but will be distorted towards those 'parts' with the largest variances;(38) we have previously termed this phenomenon **composite variable bias**.(34,35)

In theory, measuring and modelling all components offers the ideal approach to compositional data analyses. In practice, the benefits of achieving greater consistency need balancing against the demands of modelling ever more variables. As the number of components increases, there is a greater chance of violating the positivity assumption, which requires that within every stratum, there must be a non-zero probability of all (relevant) values of the exposure being observed.(39) The choice of whether to focus on the 'whole' or the 'parts' will therefore involve balancing the desired degree of consistency with the quality and availability of the data.

CONSIDERING COMPOSITE DERIVED VARIABLES

Within a DAG, a composite derived variable appears similar to a 'whole' variable in compositional data, with two or more parents causing a fully determined child (**Figure 1**). There are however some important features of composite derived variables that make them particularly challenging for causal effect estimation. First, many composite derived variables cannot be directly measured; they can only be known once the parents themselves are known.(10) Second, the parent variables may be subject to a range of functional transformations besides addition, including subtraction, division, and exponentiation.(11) Finally, the parent components may not occur at the same moment in time, giving them different temporal positions within the DAG. It is therefore extremely important to consider the nature and purpose of every composite derived variable being considered as a potential exposure or outcome.

Choosing the target estimand

Composite derived variables are commonly constructed for one of two reasons:

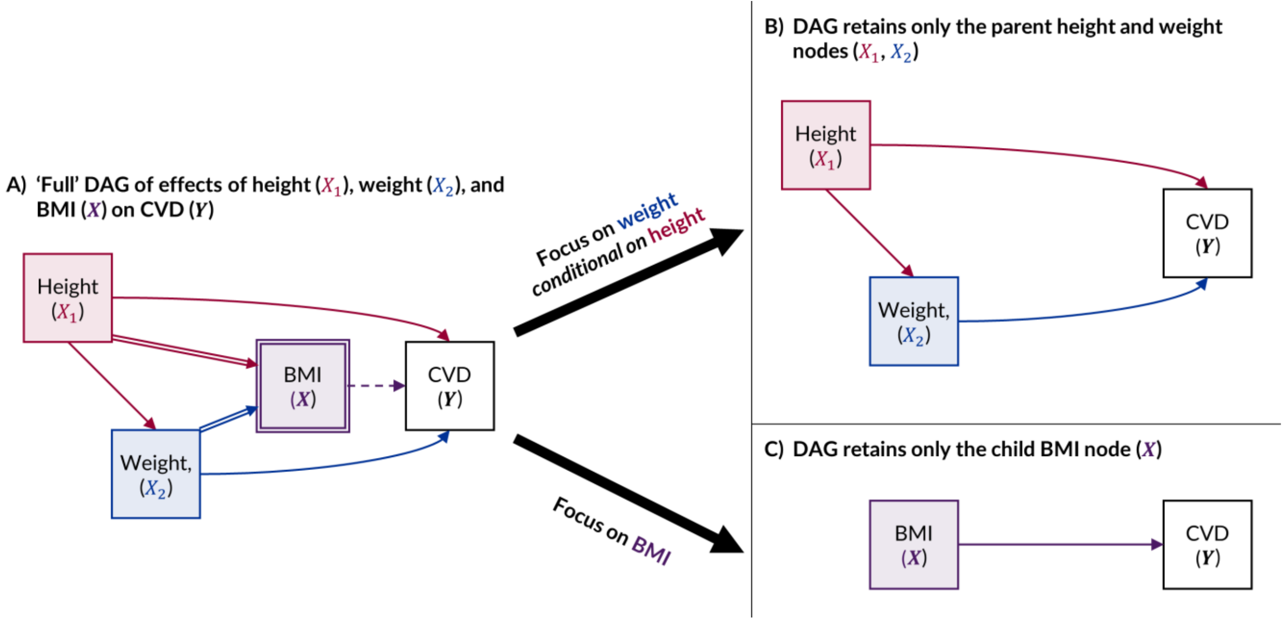
- To *summarize* several interrelated variables (e.g., deprivation index) into a single variable, either to capture a latent concept (e.g., socioeconomic circumstances), define a multifactorial state (e.g., metabolic syndrome), or provide a global summary (e.g., disease activity scores).
- To *standardize* one or more variables against one or more other variables (e.g., GDP per capita), either to account for another variable (e.g., population size), or rescale to a common unit (e.g., percentage change).

Whether a composite variable has been constructed to summarize or to standardize has immediate implications for its analysis and/or interpretation. The creation of a summary variable implies an interest in estimating the average effect of, or on, the parent variables. Conversely, composite variables that seek to standardize imply an interest in one or more parent variables while controlling for one or more 'nuisance' components. Ratio variables and change score variables, for example, both seek to isolate one parent from another, using division and subtraction, respectively.(19,22,24) Unfortunately, such approaches simply transform, rather than remove, the nuisance components.(19,22,24,34) Causal analyses of ratio variables and change score variables are hence particularly prone to inferential bias.(19,22,24,34) For most standardized composite variables, it is likely that the true target of interest is one or more target parent variables *conditional* on one or more nuisance parent variables. Rather than attempting an algebraic solution, such conditioning should be attempted using an appropriate approach, such as covariate adjustment within a linear regression model.

To illustrate, we consider the causal effect of body mass index (BMI = weight/height-squared) on the risk of cardiovascular disease (**Figure 5A**). In probabilistic terms, BMI contributes no information beyond what is captured by weight and height.(18) Deterministic Node Reduction would hence reduce BMI to a barren node that may be removed without losing information about the relationship between the remaining variables (**Figure 5B**). This explains previous assertions that "*no causal knowledge is gained by estimating a nonexistent effect of body mass index*".(18) Nevertheless, BMI may still have some utility depending on the target estimand and our reasons for creating the composite derived variable.

Since BMI is constructed by dividing weight by height-squared, it seems reasonable to assume it was conceived to standardize weight by height. However, inventor Adolphe Quetelet (1832) offers no specific motivation beyond reporting that, '*weight increases approximately with the square of the height*'.(40) Similarly, Keys *et al*. (1972), who transformed the name and prominence of the index, only appear interested in finding the best proxy measure of skinfold thickness.(41) Whether BMI is intended purely as a measure of weight standardized for height or a summary of information about both weight and height cannot therefore be known from history or algorithm, but the two perspectives carry different implications. If BMI is hypothesized as a valuable joint summary of weight and 1/height-squared then focusing on the composite measure may be reasonable, notwithstanding the issues discussed below (**Figure 5C**). Alternatively, if BMI is simply a measure of weight standardized by height, then the appropriate target would be *weight conditional on height* **(Figure 5B)**. Since the two approaches likely provide different results, determining the true estimand of interest is clearly extremely important.

Figure 5 Causal directed acyclic graphs considering the causal effect of a composite derived variable (body mass index, BMI) on an outcome (cardiovascular disease, CVD).



B) DAG retains only the parent height and weight nodes (X_1, X_2)
Height (X_1)
CVD (Y)
Weight, (X_2)
A) 'Full' DAG of effects of height (X_1), weight (X_2), and BMI (X) on CVD (Y)
Height (X_1)
BMI (X)
CVD (Y)
Weight, (X_2)
Focus on weight conditional on height
Focus on BMI
C) DAG retains only the child BMI node (X)
BMI (X)
CVD (Y)

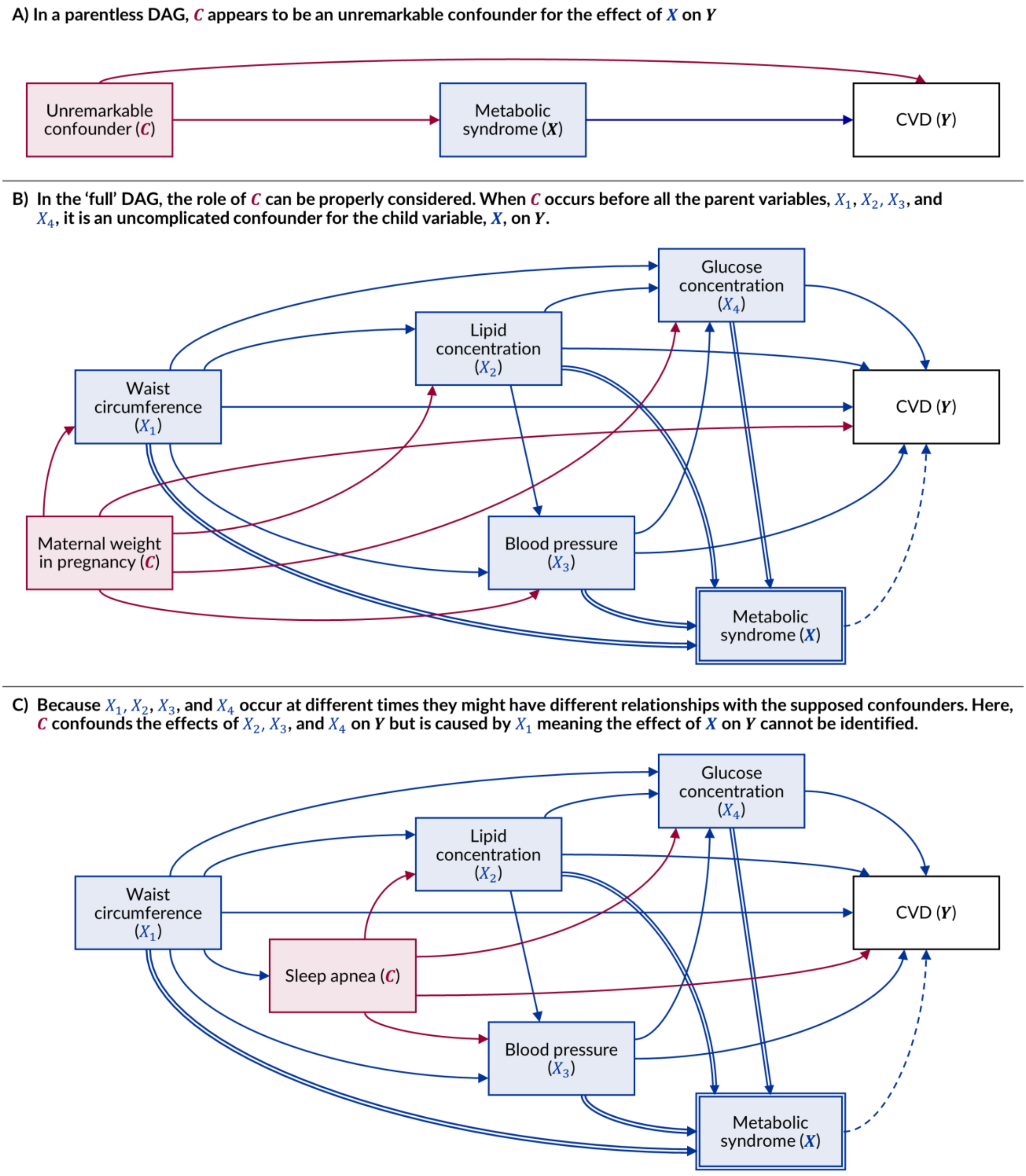
The 'full' DAG in A) shows BMI ($\boldsymbol{X}$) as a fully determined child of height (X_1) and weight (X_2). An explicit choice can be made, based on the target estimand of interest, to either retain and focus on the parent nodes X_1 and X_2, as shown in B), or to retain and focus on the child node $\boldsymbol{X}$, as shown in C).

Identifying and estimating causal effects involving composite variables

Regardless of their potential utility, most composite derived variables are likely to experience issues with satisfying the consistency and exchangeability assumptions. As with 'whole' variables in compositional data, composite derived variables have an inherent risk of consistency violations because the same value can be obtained from many different combinations of the parents. However, since composite derived variables are typically made from a more heterogeneous mix of parent variables than the 'whole' variable in compositional data, the impact of these violations may be more severe. When only the summary effect of the composite is available, the individual parent effects are lost and it becomes impossible to know which parent variables are responsible and to what extent.(42) Furthermore, due to composite variable bias, the summary effect of the composite derived variable will be skewed towards the parent variables with the largest variation within the sample,(38) which may lead to sample-specific effects that do not transport reliably.(31,43)

Regardless, the exchangeability assumption is likely the greatest barrier to identifying the causal effect of, or on, a composite derived variable. In theory, robustly identifying the causal effect of, or on, a composite derived variable requires that all confounding and selection paths are closed for all parent variables. Unfortunately, when working with the composite parent variable alone, the unique paths to and from each parent variable become conflated. Attempts to block confounding paths may therefore experience residual confounding, since only the diluted summary effect is modelled. More concerningly, if the parent variables occur at different moments in time, it is possible they will have different relationships with the supposed confounders. Indeed, it is possible that a confounder for one parent variable may be a mediator for another, leaving no means to appropriately adjust for confounding without also blocking part of the true causal effect. Identifying the causal effect of a composite derived variable, as with any analysis of multiple exposures, therefore requires no time-varying confounding.(44) To illustrate, we consider the causal effect of metabolic syndrome on the risk of cardiovascular disease (**Figure 6**). Metabolic syndrome is a composite derived variable, commonly studied as an exposure, created from waist circumference, lipid concentration, blood pressure, and glucose concentration. In a parentless DAG, we might draw the relationship between metabolic syndrome and cardiovascular disease as shown in **Figure 6A**. With metabolic syndrome depicted as a single node in time, the role of other contextual variables, represented by C, may seem unremarkable (**Figure 6A**). However, the parent variables of metabolic syndrome are unlikely to occur at the same time; instead, it is likely that some of the parents (e.g., waist circumference) may cause some of the other parents (e.g., glucose concentration). . A variable C might therefore have a very different relationship with the parent components depending on when it occurred. If C occurred before birth (e.g., maternal weight in pregnancy), then it would likely cause all the components of metabolic syndrome and be an uncomplicated confounder for its effect on the risk of cardiovascular disease (**Figure 6B**). However, if C occurred in adulthood (e.g., sleep apnea), then it might be caused by some 'earlier' parts of metabolic syndrome (e.g., waist circumference)(45), while in turn causing other 'later' parts (e.g., glucose concentration) (**Figure 6C**).(46) Although these specific examples can be debated, the 'true' C is likely to represent multiple variables, each of which may have different relationships with the individual components of metabolic syndrome. In this case, the time separation between the different parents of metabolic syndrome therefore makes the causal effect of metabolic syndrome on the risk of cardiovascular disease impossible to identify and estimate. While such issues may sometimes be avoided with repeated measures, we believe the assumption of no time-varying confounding can only by explicitly considered by depicting all parent variables within a DAG.

Figure 6 Causal directed acyclic graphs examining the challenges of identifying a causal effect for a composite derived variable (metabolic syndrome) when the parent variables are separated in time.



A) In a parentless DAG, C appears to be an unremarkable confounder for the effect of X on Y

B) In the 'full' DAG, the role of C can be properly considered. When C occurs before all the parent variables, X_1, X_2, X_3, and X_4, it is an uncomplicated confounder for the child variable, X, on Y.

C) Because X_1, X_2, X_3, and X_4 occur at different times they might have different relationships with the supposed confounders. Here, C confounds the effects of X_2, X_3, and X_4 on Y but is caused by X_1 meaning the effect of X on Y cannot be identified.

Unremarkable confounder (C)
Metabolic syndrome (X)
CVD (Y)
Glucose concentration (X_4)
Lipid concentration (X_2)
Waist circumference (X_1)
CVD (Y)
Maternal weight in pregnancy (C)
Blood pressure (X_3)
Metabolic syndrome (X)
Glucose concentration (X_4)
Lipid concentration (X_2)
Waist circumference (X_1)
CVD (Y)
Sleep apnea (C)
Blood pressure (X_3)
Metabolic syndrome (X)

*When a composite derived variable, such as metabolic syndrome (**X**), is included in a DAG without including the parents, the distinct relationship with all other contextual variables may be overlooked. In A) the contextual variable **C** appears to be an unremarkable confounder for the effect of **X** on the outcome, risk of cardiovascular disease (CVD, **Y**), because the parents of **X** have not been considered. In fact, **C** may have very different relationships with the parent variables – waist circumference (X_1), blood lipid concentration (X_2), blood pressure (X_3), and glucose concentration (X_4) – because they do not occur at the same point in time. In B) the variable **C** (e.g., maternal weight in pregnancy) occurs before all the parent components, making it an uncomplicated confounder of **X**. In C) the variable **C** (e.g., sleep apnea) occurs in between the variables that make up metabolic syndrome; it is caused by X_1 but causes X_2, X_3, and X_4 meaning it simultaneously confounds and mediates different parts of the effect of **X** on **Y** making this effect impossible to identify.*

CONCLUSION

Deterministic variables are ubiquitous in health and social science research due to the widespread use of derived variables and the common occurrence of compositional data. Unfortunately, despite repeated warnings over many decades,(19,21–24,47) the analytic and interpretational challenges of such variables remain largely underappreciated. With appropriate care and notation, we believe that DAGs can provide a novel and effective means to transform our recognition and understanding of these issues. We therefore encourage researchers to consider including deterministic variables in their DAGs when they are planning and/or interpreting analyses involving derived variables and/or compositional data.

BOX 1: GLOSSARY OF TERMS

Barren Nodes

In a directed acyclic graph, barren nodes are nodes that do not cause any other nodes.(8)

Change score variables

Change score variables (also known as difference score variables, gain score variables, or change-from-baseline variables) are a variety of composite derived variable in which an earlier measure of a time-varying variable is subtracted from a subsequent measure of that variable.(22) For example, gestational weight gain (i.e., $\Delta W = W_1 - W_0$) is a change score variable made by subtracting a pregnant person's weight at the start of pregnancy (i.e., W_0) from their weight at the end of pregnancy (i.e., W_1).

Compositional data

Compositional data, also known as comparative data, is a form of hierarchical data (formally known as a 'mereology') that contains 'part' variables (or 'meronyms') that perfectly sum to a 'whole' variable (a 'holonym') or a constant.(12) For example, the total number of children and the total number of adults (the 'part' variables) sum to the total population (the 'whole' variable). Alternatively, the total time spent physically active and the total time spent inactive (the 'part' variables), sum to a constant (total time in a day).

Compositional Data Analysis (CoDA)

The methodological area focused on the analysis of compositional data, historically with a focus on geometric transformations.(13,14)

Composite derived variables

Composite derived variables, also known simply as composite variables or compound variables, are variables that have been functionally created from two or more parent variables.(11) The value of a composite derived variable can be known with certainty once the value of all parents are known. For example, the clinical disease activity index (CDAI) is a composite derived variable created by adding together four parent variables: the total number of swollen joints, the total number tender joints, a patient-reported measure of disease severity, and a clinician-reported measure of disease severity.(48)

Composite variable bias

Composite variable bias refers to the systematic divergence between the average causal effect of a deterministic variable and the mean-weighted average causal effect of its parent variables. In general, the average effect of the deterministic variable will be distorted towards the components with the largest variance.(38) For example, consider the causal effect of metabolic syndrome – a composite derived variable created from waist circumference, lipid concentration, blood pressure, and blood glucose concentration – on the risk of cardiovascular disease. If these parent components have different variances and different causal effects on the risk of cardiovascular disease, then we can expect that the average causal effect of metabolic syndrome will differ from the mean-weighted average causal effect of the parent components.

Coarsening

Coarsening is the process of collapsing a continuous variable, or higher-order categorical variable, into a lower-order categorical variable such as a dichotomized variable. A coarsened variable has less information than its parent variable.(28) Unless the effect equivalence assumption is met, the causal effect of, or on, a coarsened variable may produce a biased estimate of the causal effect of, or on, the parent variable.(28)

D-separation

D-separation (with an uppercase 'D') is an extension of the *d*-separation criterion for identifying whether two (sets of) variables (e.g., $\boldsymbol{X}$, $\boldsymbol{Y}$) are independent conditional on a third set of variables (e.g., $\boldsymbol{Z}$) that accounts for the additional dependencies created by deterministic variables.(1) For $\boldsymbol{X}$ and $\boldsymbol{Y}$ to be independent conditional on $\boldsymbol{Z}$ (i.e., D-separated by $\boldsymbol{Z}$), then there must be no path between $\boldsymbol{X}$ and $\boldsymbol{Y}$ where: 1) all collider nodes on that path are in $\boldsymbol{Z}$ or are descended from $\boldsymbol{Z}$; and 2) all other (non-collider) nodes are outside $\boldsymbol{Z}$ and/or not functionally determined by $\boldsymbol{Z}$.(1)

Deterministic Node Reduction

Deterministic Node Reduction is an algorithm for reducing the number of nodes in a DAG. The algorithm involves identifying all deterministic variables and transferring their incoming and outgoing probabilistic arcs to their parents.(8) The resulting barren nodes may then be removed from the graph without losing information about the relationship between the remaining variables.(8)

Derived variable

Derived variables are variables that have been functionally created from one or more parent variables.(10) They include simple derived variables and composite derived variables.

Deterministic variables

Deterministic variables are variables that are functionally determined by one or more other variables, such that their value can be known with certainty once their parents are known.(8,9) Deterministic variables occur in compositional data and when derived variables are created.

Dichotomized variables

Dichotomized variables are simple derived variables, in which a continuous or multinomial parent variable has been collapsed into two categories.(29) For example, macrosomia is a dichotomized variable made from dichotomizing birthweight; birthweights of under 4500g are considered normal and weights of 4500g or more are considered macrosomic.

Effect equivalence

The effect equivalence assumption, which applies to the estimation of causal effects involving coarsened variables, requires that all versions of the exposure that have been collapsed into a single value must have the same effect.(28) This assumption is violated if a residual dose response relationship exists that is not captured by the coarsened variable.(28) Without effect equivalence, the estimated effect of the coarsened variable will be a poorly-defined weighted average of the effects of the multiple parent versions of the exposure and their frequencies in the study population.(31) For example, the estimated effect of hypertension on the risk of vascular dementia would violate the effect equivalence assumption if, among those categorized as having hypertension, those with severe hypertension had a greater risk than those with mild hypertension.

Ratio variables

Ratio variables are composite derived variables, in which one parent variable (e.g., X) is divided by second parent variable (e.g., N).(20) For example, gross domestic product (GPD) per capita is a ratio variable made by dividing total GDP by total population.

Relative compositional effects

In compositional data, a relative compositional effect is the joint effect of increasing a specified 'part' exposure while simultaneously decreasing one or more substituting 'parts' to keep the 'whole' fixed.(12,34–36) For this reason, relative compositional effects are sometimes known as substitution effects.

Simple derived variables

Simple derived variables, also known as transformed variables, are variables that have been functionally created from a single parent variable. The value of a simple derived variable can be known with certainty once the value of its parent variable is known. For example, five-year age group is a simple derived variable that is made by categorizing the continuous age variable into five-year categories (e.g., 0-4 years, 5-9 years, 10-14 years, etc.).

Tautological associations

Tautological associations, also known as spurious organic correlations and mathematical coupling, are the self-fulfilling associations that arise when a deterministic variable is analyzed in direct relation to one of its parent variables or to a sibling deterministic variable with at least one shared parent variable.(19,23,24) For example, gestational weight gain is a deterministic variable made by subtracting the weight at the end of pregnancy from the weight at the start of pregnancy; we can therefore expect gestational weight gain to share a tautological association with both weight at the start of pregnancy and weight at the end of pregnancy. Alternatively, GDP per capita and hospital bed per capita are two ratio variables with a common parent variable (total population); we can therefore expect them to share a tautological association.

Total compositional effects

In compositional data, a total compositional effect is the effect of increasing the 'whole' variable either by intervening on the 'whole' directly or through one or more specified 'part' exposures.(12,34–36) Total compositional effects are sometimes known as ‘additive effects’ because they describe the effect of ‘adding’ to the total by increasing one or more parts.

REFERENCES

1. Spirtes P, Glymour CN, Scheines R. Causation, prediction, and search. Second edition. Cambridge, Massachussetts London: The MIT Press; 2000.
2. Glymour MM. Using causal diagrams to understand common problems in social epidemiology. *Methods Soc Epidemiol.* 2006;393–428.
3. Greenland S, Pearl J, Robins JM. Causal diagrams for epidemiologic research. *Epidemiol.* 1999;10(1):37–48.
4. Hernández-Díaz S, Schisterman EF, Hernán MA. The birth weight “paradox” uncovered? *Am J Epidemiol.* 2006;164(11):1115–1120.
5. Cole SR, Platt RW, Schisterman EF, Chu H, Westreich D, Richardson D, et al. Illustrating bias due to conditioning on a collider. *Int J Epidemiol.* 2010;39(2):417–420.
6. Westreich D, Greenland S. The Table 2 Fallacy: Presenting and Interpreting Confounder and Modifier Coefficients. *Am J Epidemiol.* 2013;177(4):292–298.
7. Tennant PWG, Tomova GD, Murray EJ, Kellyn F Arnold, Matthew P Fox, Mark S Gilthorpe. Lord’s “paradox” explained: the 50-year warning on the use of “change scores” in observational data [preprint]. *arXiv*. 2023. (doi:10.48550/arXiv.2302.01822). Accessed July 25, 2023.
8. Shachter RD. Probabilistic Inference and Influence Diagrams. *Oper Res.* 1988;36(4):589–604.
9. Shachter RD. Decisions and dependence in influence diagrams. *Proceedings of the Eighth International Conference on Probabilistic Graphical Models*. 2016;52:462–473.
10. Shahar E. Metabolic syndrome? A critical look from the viewpoints of causal diagrams and statistics. *J Cardiovasc Med.* 2010;11(10):772–779.
11. Song MK, Lin FC, Ward SE, Fine JP. Composite Variables: When and How. *Nurs Res.* 2013;62(1):45–49.
12. Arnold KF, Berrie L, Tennant PWG, Gilthorpe MS. A causal inference perspective on the analysis of compositional data. *Int J Epidemiol.* 2020;49(4):1307–1313.
13. Aitchison J. The Statistical Analysis of Compositional Data. *J R Stat Soc Ser B Methodol.* 1982;44(2):139–177.
14. Filzmoser P, Hron K, Templ M. Applied Compositional Data Analysis: With Worked Examples in R. 1st ed. New York, USA: Springer Cham; 2018.
15. Pearl J. Causality: models, reasoning, and inference. 2nd ed. Cambridge, UK: Cambridge Univ. Press; 2009.
16. Lemeire J, Steenhaut K, Maes S. Causal Inference on Data Containing Deterministic Relations. 2008. http://parallel.vub.ac.be/~jan/papers/JanLemeire_CausalInferenceWithDeterministicRelations2008.pdf. Accessed July 25, 2023.
17. Geiger D, Verma T, Pearl J. Identifying independence in bayesian networks. *Networks.* 1990;20(5):507–534.
18. Shahar E. The Association of Body Mass Index With Health Outcomes: Causal, Inconsistent, or Confounded? *Am J Epidemiol.* 2009;170(8):957–958.

19. Pearson K. Mathematical contributions to the theory of evolution.—On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proc R Soc Lond.* 1897;60(359–367):489–498.

20. Kronmal RA. Spurious correlation and the fallacy of the ratio standard revisited. *J R Stat Soc Ser A Stat Soc.* 1993;156(3):379–392.

21. Oldham PD. A note on the analysis of repeated measurements of the same subjects. *J Chronic Dis.* 1962;15(10):969–977.

22. Tennant PWG, Arnold KF, Ellison GTH, Gilthorpe MS. Analyses of 'change scores' do not estimate causal effects in observational data. *Int J Epidemiol.* 2022;51(5):1604–1615.

23. Archie JP. Mathematic Coupling of Data: A Common Source of Error. *Ann Surg.* 1981;193(3):296–303.

24. Wishart J, Neyman J. Lectures and Conferences on Mathematical Statistics and Probability. *J R Stat Soc Ser Gen.* 1952;116(1):91.

25. Razak F, Subramanian S, Sarma S, Kawachi I, Berkman L, Davey Smith G, *et al.* Association between population mean and distribution of deviance in demographic surveys from 65 countries: cross sectional study. *BMJ*. 2018;k3147.

26. Choi G, Buckley JP, Kuiper JR, Keil AP. Log-transformation of Independent Variables: Must We? *Epidemiol*. 2022;33(6):843–853.

27. Vansteelandt S, Bekaert M, Claeskens G. On model selection and model misspecification in causal inference. *Stat Methods Med Res.* 2012;21(1):7–30.

28. Stitelman OM, Hubbard AE, Jewell NP. The impact of coarsening the explanatory variable of interest in making causal inferences: Implicit assumptions behind dichotomizing variables. *U.C. Berkeley Division of Biostatistics Working Paper Series.* 2010. https://biostats.bepress.com/ucbbiostat/paper264. Accessed July 25, 2023.

29. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. *Stat Med.* 2006;25(1):127–141.

30. Rehkopf DH, Glymour MM, Osypuk TL. The Consistency Assumption for Causal Inference in Social Epidemiology: When a Rose Is Not a Rose. *Curr Epidemiol Rep.* 2016;3(1):63–71.

31. Hernán MA, VanderWeele TJ. Compound treatments and transportability of causal inference. *Epidemiology*. 2011;22(3):368-377.

32. Marshall J. Coarsening bias: How coarse treatment measurement upwardly biases instrumental variable estimates. *Polit Anal.* 2016;24(2):157–171.

33. Tudball MJ, Bowden J, Hughes RA, Ly A, Munafò MR, Tilling K, Zhao Q, Davey Smith G. Mendelian randomisation with coarsened exposures. *Genet Epidemiol.* 2021;45(3):338–350.

34. Tomova GD, Arnold KF, Gilthorpe MS, Tennant PWG. Adjustment for energy intake in nutritional research: a causal inference perspective. *Am J Clin Nutr.* 2022;115(1):189–198.

35. Tomova GD, Gilthorpe MS, Tennant PWG. Theory and performance of substitution models for estimating relative causal effects in nutritional epidemiology. *Am J Clin Nutr.* 2022;116(5):1379–1388.

36. Breskin A, Murray EJ. Commentary: Compositional data call for complex interventions. *Int J Epidemiol.* 2020;49(4):1314–1315.

37. Willett W, Stampfer MJ. Total energy intake: implications for epidemiologic analyses. *Am J Epidemiol.* 1986;124(1):17–27.

38. Sharma S. Applied multivariate techniques. John Wiley & Sons, Inc.; 1995.

39. Westreich D, Cole, SR. Invited commentary: positivity in practice. *Am J Epidemiol.* 2010;171(6), 674-677.

40. Quetelet A. Recherches sur le poids de l'homme aux différens âges. *Nouv Mém Académie R Sci B-lett Brux*. 1832;7(1):1.

41. Keys A, Fidanza F, Karvonen MJ, Kimura N, Taylor HL. Indices of relative weight and obesity. *J Chronic Dis*. 1972;25(6–7):329–43.

42. McKenna SP, Heaney A. Composite outcome measurement in clinical research: the triumph of illusion over reality? *J Med Econ*. 2020;23(10):1196–204.

43. Lesko CR, Buchanan AL, Westreich D, Edwards JK, Hudgens MG, Cole SR. Generalizing study results: a potential outcomes perspective. *Epidemiol.* 2017;28(4):553.

44. Mansournia MA, Etminan M, Danaei G, Kaufman JS, Collins G. Handling time varying confounding in observational research. *BMJ*. 2017;359.

45. Gami AS, Caples SM, Somers VK. Obesity and obstructive sleep apnea. *Endocrinol Metab Clin.* 2003;32(4):869–894.

46. Botros N, Concato J, Mohsenin V, Selim B, Doctor K, Yaggi HK. Obstructive sleep apnea as a risk factor for type 2 diabetes. *Am J Med.* 2009;122(12):1122–1127.

47. Andersen B. Methodological errors in medical research: an incomplete catalogue. Oxford, UK: Blackwell scientific publications; 1993.

48. Aletaha D, Smolen J. The Simplified Disease Activity Index (SDAI) and the Clinical Disease Activity Index (CDAI): a review of their usefulness and validity in rheumatoid arthritis. *Clin Exp Rheumatol.* 2005;23(5):S100.