Commentary: the formal approach to quantitative causal inference in epidemiology: misguided or misrepresented?

Rhian M. Daniel*, Bianca L. De Stavola* and Stijn Vansteelandt†

*LSHTM Centre for Statistical Methodology
and Medical Statistics Department
London School of Hygiene and Tropical Medicine, U.K.

†Department of Applied Mathematics, Computer Sciences and Statistics
Ghent University, Belgium

To appear in: International Journal of Epidemiology

Abstract

Two recent articles, one by Vandenbroucke, Broadbent and Pearce (henceforth VBP) [1] and the other by Krieger and Davey Smith (henceforth KDS) [2], criticise what these two sets of authors characterise as the main stream of the modern "causal inference" school in epidemiology.

The criticisms made by these authors are severe; VBP label the field as both "wrong in theory" and "wrong in practice", whilst KDS—at least in some settings—feel that the field not only "bark[s] at the wrong tree" but "miss[es] the forest entirely". More specifically, the school of thought, and the concepts and methods within it, are painted as being applicable only to a very narrow range of investigations, to the exclusion of most of the important questions and study designs in modern epidemiology, such as the effects of genetic variants, the study of ethnic and gender disparities, and the use of study designs that do not closely mirror RCTs. Furthermore, the concepts and methods are painted as being potentially highly misleading even within this narrow range in which they are deemed applicable.

We believe that most of VBP's and KDS's criticisms stem from a series of misconceptions about the approach they criticise. In this response, therefore, we aim first to paint a more accurate picture of the formal causal inference approach (section 1), and then to outline the key misconceptions underlying VBP's and KDS's critiques (section 2). KDS, in particular, criticise DAGs using three examples to do so. Their discussion highlights further misconceptions concerning the role of DAGs in causal inference, and so we devote section 3 to addressing these. In our Discussion we present further objections we have to the arguments in the two papers before concluding that the clarity gained from adopting a rigorous framework is an asset, not an obstacle, to answering more reliably a very wide range of causal questions using data from observational studies of many different designs.

1 An introduction to the formal approach to quantitative causal inference in epidemiology

1.1 Labels

VBP characterise the main stream view within what they call the "causal inference movement in epidemiology" as belonging to the "restricted potential outcomes approach", which they define to be the approach in which only the effects of exposures that correspond to currently humanly feasible interventions can be studied. KDS focus instead on DAGs (rather than potential outcomes) as the main target of their criticism. However, in many places they appear to (wrongly) conflate DAGs and potential ouctomes, and they certainly share the misconception that only currently humanly feasible interventions can be studied within this approach.

As we discuss later (see misconception 1), we strongly disagree with this characterisation. We also don't much like the term 'movement', and so—for want of a better label, and to avoid cumbersome repetitive descriptions—we'll call the school of thought that both VBP and KDS have in their sight the 'Formal Approach to quantitative Causal inference in Epidemiology', or FACE. In the next sections we describe what we see as the core principles of this approach, with examples of where these have been illuminating and enabled causal analyses under less restrictive assumptions.

1.2 The core principles of the FACE

The broad features that characterise the majority of the work done by the FACE are, having first thought carefully about the precise nature of the causal question to be addressed[1], to convert this into a precise quantity to be estimated (*i.e.* a causal estimand), typically using the notation of potential outcomes. This is an obvious first step, but one that is often ignored in applied practice where researchers may jump to modelling associations and presenting their results in terms of *e.g.* odds ratios or hazard ratios, while foregoing the more interesting and concrete scientific questions *e.g.* what would the risk of this outcome be if one could eliminate the exposure? This moreover allows one to be rigorous about the assumptions (*e.g.* consistency, conditional exchangeability and positivity) under which the estimand can be identified from the data at hand, and then for flexible estimation strategies to be developed that are valid under these assumptions. Finally, tools are recommended to assess quantitatively the sensitivity of the results to plausible departures from the assumptions, to aid interpretation, and to discuss possible misinterpretation, of the results. In the Supplementary Material we give examples of causal estimands and describe the most commonly invoked assumptions needed for their identification in the context of a simplified investigation of the effect of maternal urinary tract infections during pregnancy on low birthweight.

[1]The causal question one ideally wishes to address may often be replaced by a similar causal question that can more feasibly be addressed given the constraints of the data at hand. There is a trade-off here. Noone wants 'the right answer to entirely the wrong question'; indeed, this is what has led the FACE to recommend against "retreating into the associational haven" but rather "to take the causal bull by the horns" [3]. But presumably equally uncontroversial is the observation that 'an entirely wrong answer to the right question' is also futile. Arriving at a good compromise between these two competing concerns is one of the many important tasks facing applied researchers.

1.3 The advantages of adopting this approach

In many settings (problems involving time-dependent confounding and mediation are good examples [4–9]), the increased formality characteristic of the FACE has highlighted the implausibility of the assumptions (*e.g.* no 'feedback' between exposure and confounder) required for standard analysis strategies to give meaningful answers to the causal questions being posed, and has led to improved alternatives (*e.g.* g-methods) that are increasingly widely used in practice [10–13]. The FACE has moreover given rise to an array of methods for nonlinear instrumental variable analysis [14–16] and for nonlinear mediation analysis [9, 17–23] where only *ad hoc* and biased approaches existed before.

Other examples where this approach has led to new insights and/or methods include the low birth weight and obesity "paradoxes" [24–27] (see further discussion in section 3), the comparison of dynamic regimes [28], the impact of measurement error [29, 30], noncompliance in clinical trials [31], distinguishing confounding from non-collapsibility [32], and many more.

More recently, and looking to the future, the advent of omics technologies, electronic health records and other settings that lead to high-dimensional data, means that machine learning approaches to data analysis will become increasingly important in epidemiology. For this to be a successful approach to drawing causal inferences from data, the predictive modelling aspects (to be performed by the machine) must be separated from the subject-matter considerations such as the specification of the estimand of interest, and the encoding of plausible assumptions concerning the structure of the data generating process (to be performed by humans). Whereas traditional epidemiological approaches to the analysis of data naturally blur the two aspects, the FACE makes the distinction explicit, and hence allows machine learning methods to be successfully employed [33].

1.4 An enabling or a paralysing approach?

Its emphasis on definitions and assumptions has sometimes given the false impression that the FACE is a 'paralysing' approach. How should the applied epidemiologist proceed in settings where clear definitions are hard and assumptions are violated, but nevertheless quantitative causal inference is needed? The advice that accompanies the theory is pragmatic *e.g.*

> The more precise we get the higher the risk of nonpositivity in some subsets of the study population. In practice, we need a compromise. [Hernán and Robins [34]]

and the emphasis is on adding to the statistical toolbox so that a greater range of questions can be addressed under less strict assumptions, and sensitivity analyses carried out so that appropriate transparency and scepticism enter the interpretation of results:

> Methodology almost never perfectly corresponds to the complex phenomena that give rise to our data. Methodology within a field ought to advance in expanding the range of questions that can be addressed, in relaxing the assumptions required, and in allowing investigators to assess the sensitivity of conclusions to violations in the assumptions. [VanderWeele and Vansteelandt [35]]

1.5 The focus of causal enquiries in epidemiology

We contrast two statements:

> Statement 1: Exposure E is a cause of disease D.

> Statement 2: The effect of exposure E on disease D, expressed as a risk ratio, comparing exposure level 1 vs 0, is 1.2, and this 20% increase in risk is (or is not) of sufficient magnitude to be scientifically meaningful.

Recalling the extensive discussions at the turn of this century on p-values vs. confidence intervals [36–38], the consensus among the epidemiological community—probably more so than in any other scientific community—is that knowing whether or not an exposure causes a disease (statement 1) is less important than knowing whether or not an exposure causes a disease to at least a minimally scientifically meaningful extent (statement 2). To be able to judge whether a scientifically meaningful effect is attained, it should therefore be clear from the results of an epidemiological study (i) what is the meaning of the exposure and (ii) what effect size measure is being used. For example, to understand statements such as "weight loss which was unintentional or ill-defined was associated with excess risk of 22 to 39 %" [39], one needs to understand the distribution of weight loss.

We believe that some of the apparent discrepancies between the philosophical and epidemiological standpoints on causality stem from a failure to acknowledge the difference between the two statements above, and the different levels of care and detail required when inferring such statements from data.[2]

2 Misconceptions about the FACE in VBP and KDS

There are three main shared misconceptions on which VBP and KDS build their arguments. We discuss each in turn below.

Misconception 1: The dominant view in the FACE is that hypothetical interventions must be currently humanly feasible

This idea is central to much of VBP's and KDS's criticisms of the FACE, but we do not believe it to be a correct characterisation of the dominant views within the field.

The FACE advocates having in mind hypothetical interventions that are ideally (close to being) *unambiguously defined*, and this is what is evident from the quotations chosen by VBP. We do not agree with their deduction from these quotations (nor do we interpret from the opinions expressed in the field more generally) that these hypothetical interventions need be currently humanly feasible, except of course when the purpose of the investigation is to guide imminent practical policy decisions. The statement by VBP on page 6,

> In order for an intervention to be well specified [...], it is not necessary that the intervention can be done. There is a difference between specifying and doing.

is uncontentious in our view; *sufficient specificity* is the ideal, and not feasibility.

In spite of this, the work from the FACE makes explicit that the results from a causal analysis relate to all hypothetical interventions, whether feasible and unambiguously defined or not, that—as well as the usual conditional exchangeability assumptions—satisfy the so-called consistency assumption. That is, all hypothetical interventions which are *non-invasive* in the following sense: if they

[2] It is well-known in many settings that effect estimation requires additional assumptions on top of what is required for testing the causal null hypothesis, *e.g.* methods that use instrumental variables [40].

were applied to set the exposure to some value x for all subjects, they would not change the outcome in subjects who happen to have that exposure level x, from what was actually observed.

Furthermore, since consistency at an individual level can be relaxed to a slightly weaker version of the same assumption[3], Hernán and VanderWeele [41, 42] show that it is possible to proceed even when a single non-invasive hypothetical intervention seems inconceiveable, provided that a non-invasive ensemble of hypothetical interventions is conceivable. For example, in an observational study of the effects of obesity, the work by Hernán and VanderWeele [41] shows how the interpretation of any causal effect measure estimated from a typical observational study pertains (under all other relevant assumptions) to a stochastic complex intervention that shifts the distribution of many different obesity-related exposures. Knowledge about the effects of such a hypothetical intervention is of limited value for immediate practical policy decisions, but is relevant for scientific understanding.

A growing body of work from the FACE is therefore focused on epidemiologically important exposures for which certainly no humanly feasible intervention is known, and often no single non-invasive hypothetical intervention could be conceived of for which the observational data are informative. For example, Bekaert *et al.* [43] investigate the impact of hospital-acquired infection on mortality in critically ill patients, with the aim of estimating the intensive care unit mortality risk that would have been observed had all such infections been avoided. Their analysis aims to give insight on how harmful these infections are, even though no feasible intervention exists that could prevent infection for all. By the consistency assumption, the authors view their results as being informative about the net effect of infection. This effect may differ from the effect of an intervention to prevent infection, which—if it could be designed—would likely do more than just prevent infection. Other exposures that have been recently studied in this context are, for example, socio-economic position, delirium in critically ill patients, weight change, viral clearance and depression [44–50].

Petersen and van der Laan [51] discuss the feasibility and specificity issue in a recent overview of the FACE, stating that:

> There is nothing in the structural causal model framework that requires the intervention to correspond to a feasible experiment [...] If, in addition to the causal assumptions needed for identifiability, the investigator is willing to assume that the intervention used to define the counterfactuals corresponds to a conceivable and well-defined intervention in the real world, interpretation can be further expanded to include an estimate of the impact that would be observed if that intervention were to be implemented in practice.

Much of the recent work stemming from the FACE has been dedicated to the study of mediation [9], in particular using so-called natural direct and indirect effects. These effects have been criticised by some [52] precisely because they concern hypothetical interventions that are, by their very definition, humanly unfeasible (irrespective of the variables being studied); in other words, no randomised experiment could even in principle be constructed that would allow the estimation of these effects under assumptions guaranteed to hold by design. The dominant view within the FACE is that these effects, because of the importance of the epidemiological questions they aim to address, are worthy of our attention despite the very strong unfeasibility of the hypothetical interventions they demand be imagined.

[3]The exact form of this depends on the context, but, for example, is often consistency in expectation given confounders; *i.e.* that if a hypothetical intervention were applied to set the exposure to some value x for all subjects, this would not change the conditional expectation of the outcome given confounders in subjects who happen to have that exposure level x, from the observed conditional expectation.

Misconception 2: The FACE sees the RCT as the best choice of study design for causal inference

In order to dispel this misconception, we start by proposing what we believe the characteristics of the *ideal study* to be, when inference about the total effect of a single (time-fixed) exposure is the goal. By 'ideal' we mean the study we would run if our concerns were only scientific, with no regard whatsoever for practicality, ethics or cost. We believe that such a study would have (at least) the following characteristics (and many more, of course):

1. no inclusion/exclusion criteria (so that the effect of the exposure in a variety of different groups can be separately estimated, as well as standardised effects to different (sub-)populations if relevant)

2. large sample size (also thereby ensuring large number of events if relevant)

3. an unambiguously defined set of levels for the exposure (often more than two if dose–response is of interest) allocated at random

4. long follow-up (so that short-, medium- and long-term effects can all be separately estimated)

5. rich baseline covariate data (so that effect modification can be explored)

6. no attrition, other forms of missing data, noncompliance or measurement error

It is true that point 3. says that the ideal study would be randomised (hence the fact that the FACE often talks of 'the idealised randomised experiment') but does this imply that realistic RCTs are necessarily to be viewed as better than realistic observational studies for causal inference? No; because observational studies in practice are more likely to get closer to points 1., 2., 4., and often also 5. The ideal study, which has as one of its characteristics that it is randomised, is in some respects closer to a realistic RCT, and in other ways closer to a realistic observational study. Only by knowing the specific context can a judgement be made on which is better for that context, if indeed both are feasible, ethical and practical. In many settings, when a RCT would be unfeasible, the FACE advocates having in mind the ideal (randomised) study, merely as a mental device to ensure that the observational study is designed and analysed in the most sensible fashion. This is even more valuable in complex longitudinal studies such as those that attempt to determine the optimal dynamic decision strategy [53, 54].

Since a key difference between a realistic observational study and the ideal study above is that 3. doesn't hold, a major focus of the methods arising from the FACE is how the realistic observational study can be analysed in such a way that it emulates the ideal study with respect to 3. This does not equate to the view that the FACE strives to analyse realistic observational studies in such a way that the results obtained are close to those that would have been obtained from a realistic RCT on the same exposure. The ultimate aim is to analyse realistic observational studies in such a way that the results obtained are close to those that would have been obtained from the ideal study, one feature of which is that the exposure is randomised. These two aims are different, and an investigation of this difference led to importamt insights regarding the HRT controversy by Hernán *et al.* [55].

Taken out of context, the title of the article by Hernán *et al.* "Observational studies analyzed like randomized experiments" could wrongly be taken to strengthen this misconception, that:

> Proponents of [the FACE] assume and promote the pre-eminence of the randomized controlled trial (RCT) for assessing causality; other study designs (*i.e.* observational studies) are then only considered valid and relevant to the extent that they emulate RCTs. [VBP, page 2]

On the contrary, Hernán *et al.* were not advocating that observational studies *should be* analysed like randomised experiments[4] (they dropped many years of follow-up from their data, together with many subjects who would not have met the trial's eligibility criteria, and ignored the information they had on treatment discontinuation, in order to emulate the intent-to-treat analysis performed in the RCT: it would be madness to advocate any of these measures as the best analysis of the observational data) but merely showing that if one did so, the contradiction between the results from the RCT and observational studies would be nearly eliminated, in order to challenge the dominant view at the time that the contradiction was due to unmeasured confounding in the observational studies. Incidentally, this work by Hernán *et al.* on the HRT controversy is an example of hypothesis elimination, as advocated by VBP and KDS.

As further evidence that this misconception is unfounded, we refer here to the large body of work from the FACE on the analysis of data from retrospective study designs (*e.g.* case–control studies) [58–71].

Misconception 3: The FACE believes that sex, race and genes can't be causes; furthermore (in KDS) that racism can't be a cause

Sex, race, sexism and racism as causes

This issue, particularly with respect to race, has been the source of recent controversy [72] in part in response to [73, 74]. We see this controversy ("is race a cause"?) as something of a storm in a teacup as far as epidemiology is concerned, brought about perhaps by the different focus that philosophers and epidemiologists have when it comes to causality (note that both Glymour and Glymour [72] and VBP, which has two joint lead authors, have philosophers as lead authors, and KDS also refer extensively to the philosophical literature on causality).

We refer back to Statements 1 and 2 in section 1.5. Philosophers tend to concern themselves with the meaning of statements of type 1, whereas epidemiologists are more concerned with statements of type 2, and—very importantly—whether or not it is justified to make a statement such as statement 2 from the data at hand. It would be very strange to claim that sex and race cannot be considered in place of E in statement 1. However, using them in place of E in statement 2 requires some care.

It is the dominant view within the FACE (and we agree) that asserting that "this group of Caucasians would have had a 20% lower risk of disease D had they been Afrocaribbean" is meaningful only if its readers share a near-to common understanding of what *had they been Afrocaribbean* means, and evidently this requires further details. In the counterfactual world are they to be Afrocaribbean from conception? And in what sense? Are their genes hypothetically being switched for genes that are drawn from the distribution of genes seen in Afrocaribbeans? Are they to be brought up in their biological Caucasian families, or similar Afrocaribbean families? What constitutes *similar*? Again, the consistency (and conditional exchangeability) assumption rules out many (or all) of the

[4]Note that the same lead authors have written articles with the following titles: "Randomized trials analyzed like observational studies" [56], "Observational studies analyzed like randomized trials, and vice versa" [57].

above hypothetical interventions. In order to understand which, further details must be specified, for example, whether the Afrocaribbean study participants were brought up in biological Caucasian families or not.

Why do we think that this is a storm in a teacup? Because epidemiologists are rarely interested in what would have happened to these males had they been females, nor in what would have happened to these Caucasians had they been Afrocaribbeans; rather they are interested in one of three possible things: (1) sex and race as effect modifiers, (2) describing gender and ethnic inequalities, and then in seeing what can be done to reduce them, which, as VanderWeele and Robinson show, can be done without needing to define hypothetical interventions on sex/gender/race/ethnicity, or (3) the effect of the *perception* of race and sex, *i.e.* in the effect of racism and sexism; this is what KDS talk about in their third example. None of these requires defining hypothetical interventions on sex/gender/race/ethnicity. For (3) the hypothetical intervention would be on the *perception* of race/sex, rather than on race/sex itself; see, *e.g.* [75].

We stress that the FACE is not saying that studying sex and race is not important; evidently these factors are central to many important epidemiological research questions. The 'alarm' that KDS feel[5] follows precisely from the confusion that ensues when causal inference is too informally discussed; they have misconstrued the observation made by the FACE that it is difficult to answer the question of 'what would happen if we *changed* sex/race' and that in any case we are more likely interested in one of (1), (2) or (3), above, as saying that we should not study sex and race (or even sexism and racism) at all.

With this distinction clear, it can be seen from the applied literature on investigations of ethnicity, for example, that these investigations are indeed described using associational (not causal) language, *e.g.*

> Māori and Pacific infants were twice as likely as European infants to have a mother who was obese [...] Ethnic differences in overweight were less pronounced [Howe *et al.* [76]].

The same is seen when sex/gender is studied. For example, in the recently published UK Chief Medical Officers' guidelines on safe alcohol drinking [77], gender played a key role. The committee of experts reviewed a large body of evidence on the causal effect of alcohol consumption on health outcomes, in men and women separately, and concluded that the guidelines on safe consumption limits should be the same for both genders. This was based on a study of effect modification by gender [78]. This effect modification is associational with respect to gender (but causal with respect to alcohol consumption). The pertinent question in this context did not therefore require imagining hypothetical interventions on gender.

States, including genes, as causes

VBP discuss the FACE's view of statements such a "100 000 deaths annually are attributable to obesity" and correctly characterise one of the FACE's objections to this statement as stemming from its vagueness. The statement implies something along the lines of *had there been no obesity*, there would have been 100 000 fewer deaths annually, or *were we hypothetically to eradicate obesity*, there would be 100 000 fewer deaths annually. As discussed by Hernán and Taubman [79], the words

[5]They write "One alarming feature of [the FACE] is the re-appearance of previously rebutted causal claim that 'race' [...] cannot be a 'cause' because it is not 'modifiable' " before going on to explain that it is the effect or racism, rather than the effect of race, that is of interest to them.

in italics are ambiguous. Have those who have hypothetically lost weight lost weight from their waist, or their hips, or both, and if so in what combination? Current evidence from cardiovascular epidemiology suggests that the consequences of these different possibilities would be different. Once more, the consistency assumption helps to resolve this ambiguity, but understanding its implications requires a detailed appreciation of the distribution of obesity-related exposures in the study population, as discussed by Hernán and VanderWeele [41, 42].

What is relevant to the current misconception, in particular in relation to genes as exposures, is the following characterisation of the FACE given by VBP on page 6. They extrapolate from the issue concerning obesity and conclude that under the precepts of the FACE:

> 'states' like obesity (or hypercholesterolaemia, hypertension, carrying BRCA1 or BRCA2, male gender) can no longer be seen as causes

Thus, they have concluded that the FACE believes that the causal effects of genes (along with many other things) cannot be studied. We strongly oppose this conclusion. While hypothetical interventions on BMI are too ambiguous (to imagine an obese person as not obese, there are many other changes that need also be imagined, and a myriad possibility for these) unless one elaborates further, the idea that a mutation in the BRCA1 gene inherited at meiosis could instead hypothetically not have been inherited, although currently unfeasible to implement, is sufficiently well-specified in the sense that imagining that all other inherited genes and all environmental conditions at the time of meiosis remain the same as in the actual world, would reasonably suffice for the hypothetical intervention to be non-invasive.

There are many instances in the key texts cited by VBP, KDS and beyond where the causal effects of genetic variants are discussed by the FACE [67, 69, 80–84].

3 Further misconceptions in KDS about the role of DAGs in causal inference

The description by KDS of the role played by DAGs in causal inference is counter to what is written in the key textbooks and papers in this area, and counter to what is taught in introductory courses to causal inference. We start, therefore, by clarifying the role of DAGs in causal inference, before pointing out the key misconception that underlies many of KDS's criticisms. We end this section by pointing out further errors in their discussion of the DAGs relating to their three examples.

3.1 DAGs in statistics

As used generally in statistics, DAGs are pictorial representations of conditional independences. The absence of an arrow between two nodes in a DAG is used to represent conditional independence between the two variables represented by these two nodes, conditional on the variables represented by the nodes' parents in the graph; let us call these conditional independences 'local'. The advantage of representing local conditional independences graphically is that 'global' conditional independence statements (i.e. conditional independences between two variables given sets *other than* those represented by the nodes' parents in the graph) can be deduced from the local conditional independences used to construct the graph, via an algorithm known as d-separation [85].

3.2 DAGs in causal inference

DAGs are appealing for causal inference since the causal effects of interest can be characterised in terms of specific conditional dependencies between exposure and outcome. DAGs provide insight as to which conditional dependences characterise the effect of interest by elucidating the causal structures that would render exposure and outcome conditionally dependent. Causal structures are here implied by the data-generating mechanism, which involves information on the direction of causal effects, the absence of common causes between variables, the absence of direct effects between variables, and information on the study design. Such information, which is not contained in the data but may be available from subject-matter knowledge, can be encoded in the causal DAG.

The DAGs used in causal inference can be interrogated (using d-separation) to see if, for example, a given set of variables is sufficient to adjust for confounding given the assumptions encoded in the causal DAG. The use of DAGs has thus proved very useful in this process since humans are well-known to have poor probabilistic intuition about the consequences of conditioning or adjusting. By explicitly visualising the consequences of conditioning, DAGs help to circumvent the intuitive errors that might happen when this process is attempted informally.

We stress that the DAGs used in causal inference express a priori knowledge and hypotheses. See, for example, the paper by Robins [86] in which he shows how identical data can be analysed in different ways, when guided by different causal DAGs, according to the different possible study designs, questions of interest, and subject-matter knowledge that underpin/accompany these data.

3.3 Misconceptions regarding DAGs in KDS

In the light of the above clarifications, it is now possible to address KDS's criticisms of DAGs. They point out many times that data alone are not sufficient to arrive at the DAG nor at causal inferences ("data never speak by themselves"). This is indisputable, and is precisely why DAGs are useful in causal inference: to make the assumptions based on a priori knowledge explicit, and to facilitate the translation of a priori knowledge into a suitable statistical analysis. They write that "there is no short cut for hard thinking about the biological and social realities and processes that jointly create the phenomena we epidemiologists seek to explain", and we agree. Causal DAGs don't purport to provide such a short cut; the causal DAG is the *result* of the hard thinking, not a substitute for it, and the short cut provided is via d-separation, which enters the next step, *i.e.* in helping the transition from the result of this hard thinking to a sensible statistical analysis.

Many of their criticisms are along similar lines and follow from the same underlying confusion, *e.g.* when they write "Nor can a DAG provide insight into what omitted variables might be important". We agree of course: it is the background knowledge that leads to the DAG, and not vice versa.

On page 11, KDS indicate that the world is too complicated to hope to understand all the relevant causes of the exposure in question ("One would need infinite knowledge, after all, to generate an exhaustive list…") and we, once more, agree. However, the many examples from the FACE have demonstrated that even when the DAGs are unavoidably simplistic, they do provide much insight into the biases inherent in certain statistical analyses [87].

3.4 KDS's examples

We found the discussion by KDS of their three examples rather difficult to follow, precisely since the DAGs they allude to are not drawn. This in itself points to the usefulness of DAGs for clarity

of thought and communication in these settings.

Example 1: Pellagra

In Figure 1, we have drawn a DAG capturing KDS's discussion of the pellagra example. KDS describe the two leading hypotheses as containing the same elements but with arrows "that pointed in entirely opposite directions". We don't believe this to correspond to their description nor to the plausible relationships involved. In the "germ theory", those with a high infection rate were believed to be more likely to be institutionalised, but it would not be plausible that the infection *caused* institutionalisation; rather, both would share common causes (depicted by U in our diagram) such as poverty (and hence the capitalism hypothesis is also depicted). In the remaining hypotheses they describe, there is a causal effect of institutionalisation on pellagra infection, but via different potential mediators: contaminated food, stress and vitamin B3 deficiency. All can be depicted in a DAG as we have done in Figure 1; no reversal of any arrows is involved. Of course, subject matter knowledge is needed to reach the DAG, and data analysis is then required to evaluate which are the strongest pathways, in order to determine which hypothesis (or hypotheses) is correct. The DAG in isolation is insufficient for arriving at an explanation (or for "alone wagging the causal tale"), of course, but we are unaware of claims to the contrary.

3.4.1 Example 2: Birthweight paradox

Figure 2, which is the DAG alluded to by KDS in reference to the birthweight paradox, shows that, even if we had measured and adjusted for all confounders C of smoking and infant mortality, as long as there exist unmeasured common causes U of birthweight and infant mortality, then a comparison of the mortality rates of low birthweight babies between smoking and non-smoking mothers does not have a causal interpretation. This is because stratifying on birthweight induces a correlation between smoking and U, in such a direction that it *could* explain the paradox. As VanderWeele writes in a recent review article on this issue [88]:

> The intuition behind this explanation is that low birthweight might be due to a number of causes: one of these might be maternal smoking, another might be instances of malnutrition or a birth defect. If we consider the low birthweight infants whose mothers smoke, then it is likely that smoking is the cause of low birthweight. If we consider the low birthweight infants whose mothers do not smoke, then we know maternal smoking is ruled out as a cause for low birthweight, so that there must have been some other cause, possibly something such as malnutrition or a birth defect, the consequences of which for infant mortality are much worse. By not controlling for the common causes (U) of low birthweight and infant mortality, we are essentially setting up an unfair comparison between the smoking and non-smoking mothers. If we could control for such common causes, the paradoxical associations might go away.

VanderWeele chooses malnutrition and birth defects as possible Us, whereas KDS choose "harms during fetal development unrelated to and much worse than those imposed by smoking, *e.g.* stochastic semi-disasters that knock down birthweight, as a result of random genetic or epigenetic abnormalities affecting the sperm or egg prior to conception or arising during fertilization and embryogenesis". Is this not just a biologically more detailed description of the sort of phenomenon involved in the development of a birth defect, in which malnutrition could also play a part? In other words, the

'DAG explanation' and KDS's explanation are almost the same, and indeed, since the 'DAG explanation' only posits that such a U may exist, it subsumes KDS's more specific explanation. We don't understand their claim, therefore, that the former explanation is incorrect and misleading, while the latter is "lovely".

Their comment that, having identified the potential for collider bias in a DAG, "it is another matter entirely, however, to elucidate, empirically, if the hypothesized biases do indeed exist and if they are sufficient to generate the observed associations" is of course entirely uncontentious. This is precisely why, having identified the possibility that the paradox could be explained in this way, the FACE went on to evaluate whether or not plausible magnitudes for the effects of such U on birthweight and infant mortality would suffice to explain the reported paradoxical associations [25, 89, 90].

In summary, DAGs are neither the beginning (they arise from subject matter knowledge) nor the end (they guide the subsequent data analysis and/or sensitivity analyses), but neither has the FACE made claims to this effect.

3.4.2 Example 3: Racism

As we discussed under Misconception 3 above, KDS are in agreement with the FACE in their discussion of their third example, since hypothetical interventions on *racism* don't suffer from any of the specification problems that accompany hypothetical interventions on race discussed above and in the literature they criticise. Rather than saying that the FACE is "bark[ing] at the wrong tree, and indeed miss[ing] the forest entirely", KDS should surely aim this criticism at their fellow-critics of the FACE, *e.g.* VBP, who are the ones advocating studying the causal effects of race and sex; the FACE has merely outlined the difficulties in doing so, and entirely agrees that it is unlikely to be the true question of interest.

4 Discussion

4.1 Formality and non-invasive hypothetical interventions

In view of the difficulties of making causal enquiries based on observational data, epidemiologists have historically tended to speak only of associations. VBP rightly say that the FACE has been a response to this 'retreat to the associational haven'. While prudence is imperative, incidentally, this 'retreat' has tended to result in a lack of prudence in data analysis. Indeed, since essentially all statistical analyses are designed to measure associations, adjusted or not, the lack of a formal framework makes it impossible to distinguish clearly between analysis strategies that target the envisaged causal enquiry from those that do not. The unfortunate result has been reflected in analysis strategies that tend to induce bias, even in the ideal setting where all relevant confounding variables are perfectly measured.

To be able to identify, from across the many possible associations between exposure and outcome that one could measure, the one that targets the causal enquiry at stake, the FACE has adopted the notion of hypothetical interventions. Using such hypothetical interventions, effect measures of interest can be clearly expressed, identifying assumptions can be explicated and analysis strategies developed that are valid when these assumptions are met. The FACE thus merely aims to provide

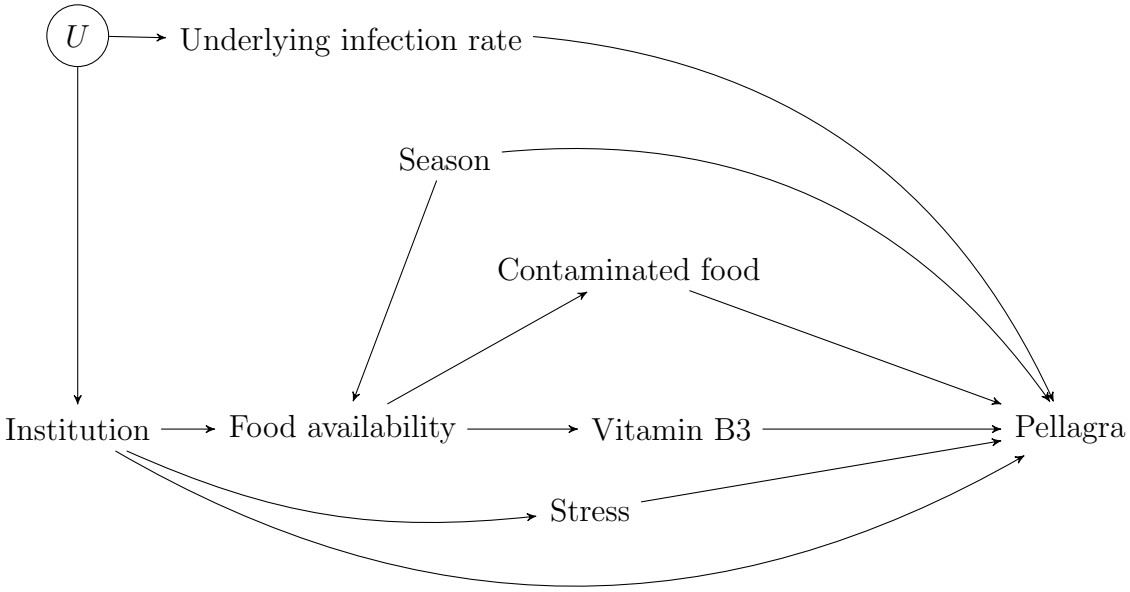

Figure 1: A causal DAG representing all the hypotheses discussed by KDS in relation to the effect of institutionalisation on pellagra infection

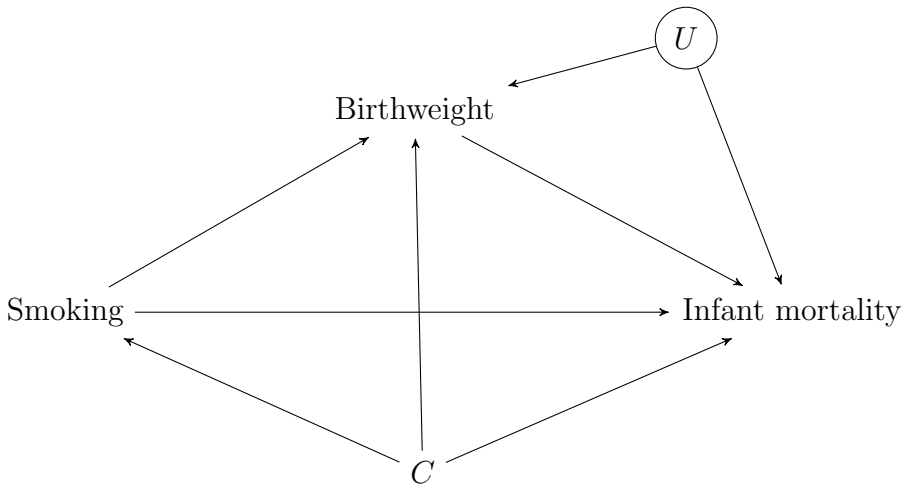

Figure 2: A causal DAG for the "birthweight paradox"

a principled framework under which causal enquiries can be approached. It does not eschew the many sources of epidemiologic information, such as time-trend data, retrospective designs, negative controls, *etc.* but rather aims to understand under what conditions such information enables causal enquiries to be answered; for examples of this work by the FACE in relation to time-trend data and negative controls, see [91–97]. In addition, it aims to caution epidemiologists that a good understanding of reported effect requires a specific understanding of the exposure and considered effect measure.

While adopting the specific interventionist framework as a philosophy, we have argued that the formality that underlies the FACE does not require the existence of humanly feasible interventions, as it targets 'non-invasive interventions' in the sense implied by the consistency assumption. We believe that many epidemiologic enquiries, except those that aim to evaluate the impact of public health interventions, implicitly have such interventions in mind.

4.2 Alternative frameworks

A number of causal theories have attempted to move away from the main stream approach as described above, by not using potential outcomes [99–101]. Some of these, in particular the decision-theoretic framework, have been useful in highlighting some strong assumptions entailed in approaches based on potential outcomes, particularly when joint or nested counterfactuals are involved. The decision-theoretic framework adheres to the same principles (one might argue even more strongly) of clearly expressing the causal target of estimation, and the assumptions under which this can be identified. Indeed, as such, in terms of data analysis, the decision-theoretic approach reproduces existing results from the potential outcomes approach, and we view it as a part of the FACE. Other causal theories, in their attempt to avoid potential outcomes, have tended to be less explicit, thereby obscuring and eventually ignoring certain selection biases. VBP and KDS similarly recommend that other philosophical frameworks for causality be adopted in epidemiology. We hope that their alternatives, which are not sufficiently specific to be fully-evaluated, will not run into the same difficulties.

Both VBP and KDS mention the need for the synthesis of evidence across multiple studies and settings. We agree with this, and view the concepts and methods of the FACE as aiding rather than impeding this endeavour, in two ways: (i) more reliable causal analyses of the individual studies contributing to a synthesis improves the reliability of the synthesised conclusion, and (ii) by being clear what question is being addressed, and under what assumptions the analysis strategy used can be deemed successful, evidence from different studies can be more reliably combined. For a recent example of where a meta-analysis came to suspect conclusions based on shortcomings in both these aspects, see [102].

VBP and KDS suggest the analysis of time-trend data, the use of negative controls, and the elimination of alternative hypotheses, but as we have discussed, these are already done within the FACE [91–97]. Arguably, the vast section of the FACE literature dedicated to sensitivity analyses has at its core the elimination (or at least consideration or evaluation) of alternative hypotheses. A novel approach to the elimination of alternative hypotheses is described by Rosenbaum [98].

VBP also imply that Pearl's framework (specifically, non-parametric structural equation models (NPSEM) [85]) is more amenable to epidemiological enquiries. While of course we view the NPSEM framework as belonging to the FACE, it is well-known that the NPSEM framework is *more* demanding in terms of the assumptions it makes than alternative frameworks within the FACE [103]. These

are specifically assumptions similar to consistency. Instead of making the consistency assumption only with respect to hypothetical interventions on the exposure, the NPSEM assumptions imply consistency with respect to hypothetical interventions on every variable in the causal diagram. We fail to follow therefore why VBP might be prepared to accept this more restrictive sub-framework, whilst viewing the larger framework that contains it as too restrictive.

4.3 Historical success stories

Both VBP and KDS draw attention to a few historic examples from epidemiology's past in which successful causal inferences were achieved without the formality advocated by the FACE. We should be cautious of basing future strategy on these 'cherry picked' success stories, without mentioning the numerous failures.

Indeed, a similar reasoning would lead one to conclude that science does not need a formal deductive theory at all, since there are obviously many examples, *e.g.* in prehistoric times, where science and knowledge acquisition progressed without formal theories. The logical error in this reasoning is that no consideration is given to the many examples where plain intuition and informal deduction have been misleading. This does not mean that informal approaches have no value; they should and do guide the design of studies and statistical analysis, but objective science eventually calls for a formal theory and approach.

We view the FACE as precisely offering formal tools to investigate cause–effect relationships. They are always guided by what KDS call IBE (inference to the best explanation). Indeed, IBE is often how one comes to investigate the specific cause–effect relationship in the first place. Given how associations can be distorted in complicated ways due to implicit/explicit conditioning or not conditioning, and how intuition, *e.g.* in mediation analysis and instrumental variable methods, breaks down as soon as non-linear relationships are at play, there is no question in our opinion, that a formal theory is needed to guide data analysis.

4.4 Concluding thoughts

Throughout its history, aspects of the FACE have been misconceived by some. Its tendency to be explicit about assumptions has often been misunderstood as if this framework needs more assumptions than traditional alternatives. This has then led people to use 'associational analyses' instead, the conclusions from which they eventually interpret causally, where causal interpretation is only justified under even stronger assumptions.

These papers by VBP and KDS highlight further misconceptions, which, if true, would mean that many important exposures would be excluded from being studied within the FACE framework and many tools, such as causal DAGs, rejected as misleading. In this response, we have attempted to correct these misconceptions, and, while stressing the clarity that comes from having a rigorous framework based on clear definitions and assumptions, we have highlighted the pragmatic considerations that should and do accompany the theory when applied in practice, together with the central role played by subject-matter knowledge. We are glad to learn about these concerns, and to be able to clarify that the FACE does not refute epidemiological questions that cannot be linked to humanly feasible interventions, nor epidemiological designs that cannot emulate aspects of randomised studies, and nor does it claim that graphical or statistical methods lessen the importance of subject-matter knowledge. Rather, the FACE aims to provide insight on what can be learned

about these questions and from these designs under the most plausible assumptions possible given the data, design and subject-matter knowledge at hand.

As Hernán [104] concluded in a recent debate on similar issues, relating to whether or not left-truncated data can meaningfully be used in causal inference:

> Exceptions to this synchronizing of the start of follow-up and the treatment strategies may be considered when the only available data (or the only data that we can afford) are left truncated. If we believe that analyzing those data will improve the existing evidence for decision-making, we must defend the use of left-truncated data explicitly, rather than defaulting into using the data without any justification.

We understand from this, and agree, that no data, and no questions are 'off limits' as long as the data are informative about the question. The core theme of the FACE is that formality allows one to assess to what extent the data at hand are informative about a particular question given subject-matter knowledge. A rejection of this framework in favour of an alternative would either mean that the new framework could do away with the need to link the data to the question, or that the required link would remain, but in an obscured and less explicit fashion. The former would be miraculous, and the latter would increase the risk of confusion and misinterpretation.

Acknowledgements

We are grateful to Jonathan Bartlett, Alex Broadbent, Karla Diaz Ordaz, Isabel dos Santos Silva, Oliver Dukes, Sander Greenland, Miguel Hernán, Dave Leon, Jamie Robins, Elizabeth Williamson, Jan Vandenbroucke and Tyler VanderWeele for stimulating discussions on these issues and/or comments on an earlier draft.

Rhian Daniel is supported by a Sir Henry Dale Fellowship jointly funded by the Wellcome Trust and the Royal Society (Grant Number 107617/Z/15/Z). The LSHTM Centre for Statistical Methodology is supported by the Wellcome Trust Institutional Strategic Support Fund, 097834/Z/11/B. Stijn Vansteelandt acknowledges support from IAP research network grant no. P07/06 from the Belgian government (Belgian Science Policy).

References

[1] Vandenbroucke JP, Broadbent A, Pearce N (2016) Causality and causal inference in epidemiology: the need for a pluralistic approach. Int J Epidemiol, Epub ahead of print.

[2] Krieger N, Davey Smith G (in press) The tale wagged by the DAG: broadening the scope of causal inference and explanation for epidemiology. Int J Epidemiol, to appear.

[3] Hernán MA (2005) Invited commentary: hypothetical interventions to define causal effects—afterthought or prerequisite? Am J Epidemiol 162:618–20.

[4] Robins JM (1986) A new approach to causal inference in mortality studies with sustained exposure periods – Application to control of the healthy worker survivor effect. Mathematical Modelling 7:1393–512.

[5] Robins JM, Hernán M, Brumback B (2000) Marginal structural models and causal inference in epidemiology. Epidemiology 11(5):550–60.

[6] Robins JM, Hernán MA (2009) Estimation of the causal effects of time-varying exposures. In: Longitudinal Data Analysis, Fitzmaurice G, Davidian M, Verbeke G, Molenberghs G, eds. New York: Chapman and Hall/CRC Press.

[7] Daniel RM, Cousens SN, De Stavola BL, Kenward MG, Sterne JAC (2013) Methods for dealing with time-dependent confounding. Stat Med 32(9):1584–618.

[8] Cole SR, Hernán MA (2002) Fallibility in estimating direct effects. Int J Epidemiol 31(1):163–5.

[9] VanderWeele T (2015) Explanation in causal inference: methods for mediation and interaction. New York, NY: Oxford University Press.

[10] Kim C, Feldman HI, Joffe M, Tenhave T, Boston R, Apter AJ (2005) Influences of earlier adherence and symptoms on current symptoms: A marginal structural models analysis. The Journal of Allergy and Clinical Immunology 115(4):810–4.

[11] Moore K, Neugebauer R, Lurmann F, Hall J, Brajer V, Alcorn S, Tager I (2008) Ambient ozone concentrations cause increased hospitalizations for asthma in children: an 18-year study in Southern California. Environmental Health Perspectives 116(8):1063.

[12] Morrison CS, Pai-Lien CHEN, Cynthia KWOK., Richardson BA, Chipato T, Mugerwa R, Byamugisha J, Padian N, Celentano DD, Salata RA (2010) Hormonal contraception and HIV acquisition: reanalysis using marginal structural modeling. AIDS 24(11):1778.

[13] Banack HR, Kaufman JS (2016) Estimating the time-varying joint effects of obesity and smoking on all-cause mortality using marginal structural models Am J Epidemiol 183(2):122–9.

[14] Vansteelandt S, Goetghebeur E (2003) Causal inference with generalized structural mean models. J R Stat Soc Series B 65:817–35.

[15] Robins JM, Rotnitzky A (2004) Estimation of treatment effects in randomized trials with noncompliance and a dichotomous outcome using structural mean models. Biometrika 91:763–83.

[16] Tchetgen Tchetgen EJ, Walter S, Vansteelandt S, Martinussen T, Glymour M (2015) Instrumental variable estimation in a survival context. Epidemiology 26(3):402–10.

[17] Robins JM, Greenland S (1992) Identifiability and exchangeability for direct and indirect effects. Epidemiology 1:143–55.

[18] Pearl J (2001) Direct and indirect effects. Proceedings of the seventeenth conference on uncertainty in artificial intelligence 411–420.

[19] VanderWeele T, Vansteelandt S (2009) Conceptual issues concerning mediation, interventions and composition. Statistics and its Interface 2:457–68.

[20] VanderWeele TJ (2009) Marginal structural models for the estimation of direct and indirect effects. Epidemiology 20(1):18–26.

[21] VanderWeele TJ, Vansteelandt S (2010) Odds ratios for mediation analysis for a dichotomous outcome. Am J Epidemiol 172(12):1339–48.

[22] Loeys T, Moerkerke B, De Smet O, Buysse A, Steen J, Vansteelandt S (2013) Flexible mediation analysis in the presence of nonlinear relations: beyond the mediation formula. Multivariate Behavioral Research. 48(6):871–94.

[23] De Stavola BL, Daniel RM, Ploubidis GB, Micali N. Mediation analysis with intermediate confounding: structural equation modeling viewed through the causal inference lens. Am J Epidemiol 181(1):64–80.

[24] Hernández-Díaz S, Schisterman EF, Hernán MA (2006) The birth weight "paradox" uncovered? Am J Epidemiol 164(11):1115–20.

[25] Whitcomb BW, Schisterman EF, Perkins NJ, Platt RW (2009) Quantification of collider-stratification bias and the birthweight paradox. Paediatric and perinatal epidemiology 23(5):394–402.

[26] Banack HR, Kaufman JS (2013) The "obesity paradox" explained. Epidemiology 24(3):461–2.

[27] Preston SH, Stokes A (2014) Obesity paradox: conditioning on disease enhances biases in estimating the mortality risks of obesity. Epidemiology 25(3):454–61.

[28] Chakraborty B, Moodie EE (2013) Statistical methods for dynamic treatment regimes. Springer.

[29] Hernán MA, Cole SR (2009) Invited Commentary: Causal diagrams and measurement bias. Am J Epidemiol 170(8):959–62.

[30] VanderWeele TJ, Hernán MA (2012) Results on differential and dependent measurement error of the exposure and the outcome using signed directed acyclic graphs. Am J Epidemiol 175(12):1303–10.

[31] Robins JM, Tsiatis AA (1991) Correcting for non-compliance in randomized trials using rank preserving structural failure time models. Communications in Statistics — Theory and Methods 20(8):2609–31.

[32] Greenland S, Robins JM, Pearl J (1999) Confounding and collapsibility in causal inference. Statistical Science 1:29–46.

[33] van der Laan MJ, Rose S (2011) Targeted Learning: Causal Inference for Observational and Experimental Data. Springer.

[34] Hernán MA, Robins JM (2016) Causal Inference. Boca Raton: Chapman & Hall/CRC, forthcoming.

[35] VanderWeele TJ, Vansteelandt S (2010) VanderWeele and Vansteelandt respond to "Decomposing with a lot of supposing" and "Mediation". Am J Epidemiol 172 (12): 1355–6.

[36] Lang JM, Rothman KJ, Cann CI (1998) That confounded P-value. Epidemiology 9:7–8.

[37] Sterne JAC, Davey Smith G (2001) Sifting the evidence—what's wrong with significance tests? BMJ 322:226–31.

[38] Poole C (2001) Low P-values or narrow confidence intervals: which are more durable? Epidemiology 12(3):291–4.

[39] Harrington M, Gibson S, Cottrell RC (2009) A review and meta-analysis of the effect of weight loss on all-cause mortality risk. Nutr Res Rev 22(1):93–108.

[40] Hernán MA, Robins JM (2006) Instruments for causal inference: an epidemiologist's dream? Epidemiology 17(4):360–72.

[41] Hernán MA, VanderWeele TJ (2011) Compound treatments and transportability of causal inference. Epidemiology 22(3):368–77.

[42] VanderWeele TJ, Hernán MA (2013) Causal inference under multiple versions of treatment. Journal of Causal Inference 1:1–20.

[43] Bekaert M, Timsit JF, Vansteelandt S, Depuydt P, Vsin A, Garrouste-Orgeas M, Decruyenaere J, Clec'h C, Azoulay E, Benoit D, Outcomerea Study Group (2011) Attributable mortality of ventilator-associated pneumonia: a reappraisal using causal analysis. Am J Respir Crit Care Med 184(10):1133–9.

[44] Schnitzer ME, Moodie EEM, van der Laan MJ, Platt RW, Klein MB (2014) Modeling the impact of Hepatitis C viral clearance on end-stage liver disease in an HIV co-infected cohort with targeted maximum likelihood estimation. Biometrics 70(1)144–52.

[45] Klein Klouwenberg PMC, Zaal IJ, Spitoni C, Ong DSY, van der Kooi AW, Bonten MJM, Slooter AJC, Cremer OL (2014) The attributable mortality of delirium in critically ill patients: prospective cohort study. BMJ 349:g6652.

[46] Huang JY, Gavin AR, Richardson TS, Rowhani-Rahbar A, Siscovick DS, Enquobahrie DA (2015) Are early-life socioeconomic conditions directly related to birth outcomes? Grandmaternal education, grandchild birth weight, and associated bias analyses. Am J Epidemiol 182(7):568–78.

[47] Rao SK, Mejia GC, Roberts-Thomson K, Logan RM, Kamath V, Kulkarni M, Mittinty MN (2015) Estimating the effect of childhood socioeconomic disadvantage on oral cancer in India using marginal structural models. Epidemiology 26(4):509–517.

[48] Cao B (2015) Estimating the effects of obesity and weight change on mortality using a dynamic causal model. PLOS ONE 10(6):e0129946.

[49] Gilsanz P, Walter S, Tchetgen Tchetgen EJ, Patton KK, Moon JR, Capistrant BD, Marden JR, Kubzansky LD, Kawachi I, Glymour MM (2015) Changes in depressive symptoms and incidence of first stroke among middle-aged and older US adults. J Am Heart Assoc 4(5)e001923.

[50] Maika A, Mittinty MN, Brinkman S, Lynch J (2014) Effect on child cognitive function of increasing household expenditure in Indonesia: application of a marginal structural model and simulation of a cash transfer programme. Int J Epidemiol 44(1):218–28.

[51] Petersen ML, van der Laan MJ (2014) Causal models and learning from data: Integrating causal modeling and statistical estimation. Epidemiology 25(3):418–26.

[52] Naimi AI, Kaufman JS, MacLehose RF (2014) Mediation misgivings: ambiguous clinical and public health interpretations of natural direct and indirect effects. Int J Epidemiol 43(5):1656–61.

[53] Didelez V (2016) Commentary: Should the analysis of observational data always be preceded by specifying a target experimental trial? Int J Epidemiol, Epub ahead of print.

[54] Hernán MA (2016) Using big data to emulate a target trial when a randomized trial is not available Am J Epidemiol, Epub ahead of print.

[55] Hernán MA, Alonso A, Logan R, Grodstein F, Michels KB, Stampfer MJ, Willett WC, Manson JE, Robins JM (2008) Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. Epidemiology 19(6):766.

[56] Hernán MA, Hernández-Díaz S, Robins JM (2013) Randomized trials analyzed like observational studies. Annals of Internal Medicine 159(8): 560–562.

[57] Hernán MA, Robins JM (2016) Observational studies analyzed like randomized trials, and vice versa. In: Gatsonis C, Morton S, eds. Methods in Comparative Effectiveness Research. Chapman & Hall/CRC Press: Boca Raton, FL.

[58] Hernán MA, Hernández-Díaz S, Robins JM (2004) A structural approach to selection bias. Epidemiology 15(5):615–25.

[59] Hudson JI, Javaras KN, Laird NM, VanderWeele TJ, Pope Jr HG, Hernán MA (2008) A structural approach to the familial coaggregation of disorders. Epidemiology 19(3):431–9.

[60] Vansteelandt S (2009) Estimating direct effects in cohort and case-control studies. Epidemiology 20(6):851–60.

[61] Tchetgen Tchetgen EJ, Robins J (2010) The semiparametric case-only estimator. Biometrics 66(4):1138–44.

[62] Didelez V, Kreiner S, Keiding N (2010) Graphical models for inference under outcome-dependent sampling. Statistical Science 25(3):368–87.

[63] Kuroki M, Cai Z, Geng Z (2010) Sharp bounds on causal effects in case-control and cohort studies. Biometrika 97(1):123–32.

[64] VanderWeele TJ, Vansteelandt S (2011) A weighting approach to causal effects and additive interaction in case-control studies: marginal structural linear odds models. Am J Epidemiol 174(10):1197–203.

[65] Bowden J, Vansteelandt S (2011) Mendelian randomization analysis of case-control data using structural mean models. Stat Med 30(6):678–94.

[66] Tchetgen Tchetgen EJ, Rotnitzky A (2011) Double-robust estimation of an exposure-outcome odds ratio adjusting for confounding in cohort and case-control studies. Stat Med 30(4):335–47.

[67] Vansteelandt S, Lange C (2012) Causation and causal inference for genetic effects. Human Genetics 131(10):1665–76.

[68] Berzuini C, Vansteelandt S, Foco L, Pastorino R, Bernardinelli L (2012) Direct genetic effects and their estimation from matched case-control data. Genetic Epidemiology 36(6):652–62.

[69] VanderWeele TJ, Asomaning K, Tchetgen EJ, Han Y, Spitz MR, Shete S, Wu X, Gaborieau V, Wang Y, McLaughlin J, Hung RJ (2012) Genetic variants on 15q25. 1, smoking, and lung cancer: an assessment of mediation and interaction. Am J Epidemiol 175(10):1013–20.

[70] Persson E, Waernbaum I (2013) Estimating a marginal causal odds ratio in a case-control design: analyzing the effect of low birth weight on the risk of type 1 diabetes mellitus. Stat Med 32(14):2500–12.

[71] Kennedy EH, Sjolander A, Small DS (2015) Semiparametric causal inference in matched cohort studies. Biometrika 102(3):739–46.

[72] Glymour C, Glymour MR (2014) Commentary: race and sex are causes. Epidemiology 25(4):488–90.

[73] VanderWeele TJ, Hernán MA. (2012) Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures, with application to the effects of race and sex. In: Causality: Statistical Perspectives and Applications, Berzuini C, Dawid AP, Bernardinelli L. John Wiley & Sons.

[74] VanderWeele TJ, Robinson WR (2014) On causal interpretation of race in regressions adjusting for confounding and mediating variables. Epidemiology 25(4):473.

[75] Bertrand SM, Chung H, Fern A, Anne M (2004) Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination. In: The American Economic Review.

[76] Howe LD, Ellison-Loschmann L, Pearce N, Douwes J, Jeffreys M, Firestone R (2015) Ethnic differences in risk factors for obesity in New Zealand infants. Journal of Epidemiology and Community Health 69(6):516–22.

[77] UK Chief Medical Officers (2016) Alcohol Guidelines Review — Report from the Guidelines development group to the UK Chief Medical Officers. `https://www.gov.uk/government/consultations/health-risks-from-alcohol-new-guidelines`.

[78] VanderWeele TJ (2009) On the distinction between interaction and effect modification. Epidemiology 20(6):863–71.

[79] Hernán MA, Taubman SL (2008) Does obesity shorten life? The importance of well-defined interventions to answer causal questions. Int J Obes 32(3):S8–14.

[80] VanderWeele TJ, Hernández-Díaz S, Hernán MA (2010) Case-only gene-environment interaction studies: when does association imply mechanistic interaction? Genetic Epidemiology 34:327–334.

[81] VanderWeele TJ, Laird NM (2011) Tests for compositional epistasis under single interaction-parameter models. Annals of Human Genetics 75:146–156.

[82] Chen J, Kang G, VanderWeele TJ, Zhang C, Mukherjee B (2012) Efficient designs of gene-environment interaction studies: implications of Hardy-Weinberg equilibrium and gene-environment independence. Stat Med 31:2516–30.

[83] VanderWeele TJ. Ko Y-A, Mukherjee, B (2013) Environmental confounding in gene-environment interaction studies. Am J Epidemiol 178:144–152.

[84] Joshi AD, Lindstrom S, Husing A, Barrdahl M, VanderWeele TJ, Campa D, Canzian F, Gaudet MM, Figueroa JD, Baglietto L, Berg CD, Buring JE, Chanock SJ, Chirlaque MD, Diver WR, Dossus L, Giles GG, Haiman CA, Hankinson SE, Henderson BE, Hoover RN, Hunter DJ, Isaacs C, Kaaks R, Kolonel LN, Krogh V, Le Marchand L, Lee IM, Lund E, McCarty CA, Overvad K, Peeters PH, Riboli E, Schumacher F, Severi G, Stram DO, Sund M, Thun MJ, Travis RC, Trichopoulos D, Willett WC, Zhang S, Ziegler RG, Kraft P; Breast and Prostate Cancer Cohort Consortium (BPC3) (2014) Additive interactions between GWAS-identified susceptibility SNPs and breast cancer risk factors in the BPC3. Am J Epidemiol 180:1018–27.

[85] Pearl J (2009) Causality. Cambridge University Press.

[86] Robins JM (2001) Data, design, and background knowledge in etiologic inference. Epidemiology 12(3):313–20.

[87] Glymour MM (2006) Using causal diagrams to understand common problems in social epidemiology. Methods in social epidemiology, 393–428.

[88] VanderWeele TJ (2014) Commentary: Resolutions of the birthweight paradox: competing explanations and analytical insights. Int J Epidemiol 43(5):1368–73.

[89] Basso O, Wilcox AJ (2009) Intersecting birth weight-specific mortality curves: solving the riddle. Am J Epidemiol 169(7):787–797.

[90] VanderWeele TJ, Mumford SL, Schisterman EF (2012) Conditioning on intermediates in perinatal epidemiology. Epidemiology 23:1–9.

[91] Bor J, Moscoe E, Mutevedzi P, Newell ML, Brnighausen T (2014) Regression discontinuity designs in epidemiology: causal inference without randomized trials. Epidemiology 25(5):729–37.

[92] Geneletti S, O'Keeffe AG, Sharples LD, Richardson S, Baio G (2015) Bayesian regression discontinuity designs: Incorporating clinical knowledge in the causal analysis of primary care data. Stat Med 34(15):2334–52.

[93] Bor J, Moscoe E, Baernighausen T (2015) Three approaches to causal inference in regression discontinuity designs. Epidemiology 26(2):E28–E30.

[94] Lipsitch M, Tchetgen Tchetgen E, Cohen T (2010) Negative controls: a tool for detecting confounding and bias in observational studies. Epidemiology 21:383–88.

[95] Abadie A (2005) Semiparametric difference-in-differences estimators. The Review of Economic Studies 72(1):1–19.

[96] Athey S, Imbens GW (2006) Identification and inference in nonlinear difference-in-differences models. Econometrica 74(2):431–97.

[97] Sofer T, Richardson DB, Colincino E, Schwartz J, Tchetgen Tchetgen EJ (2015) On simple relations between difference-in-differences and negative outcome control of unobserved confounding. Harvard University Biostatistics Working Paper Series. Working Paper 194. `http://biostats.bepress.com/harvardbiostat/paper194`.

[98] Rosenbaum PR (2015) Some counterclaims undermine themselves in observational studies. Journal of the American Statistical Association 110(512):1389–98.

[99] Dawid P (2012) The decision theoretic approach to causal inference. In: Causality: Statistical Perspectives and Applications, Berzuini C, Dawid AP, Bernardinelli L. John Wiley & Sons.

[100] Aalen OO (1987) Dynamic modelling and causality. Scandinavian Actuarial Journal. 1:177–90.

[101] Commenges D, Gégout-Petit A (2009) A general dynamical model with causal interpretation. JRRS-B 71(4):1–18.

[102] Kalkhoran S, Glantz SA (2016) E-cigarettes and smoking cessation in real-world and clinical settings: a systematic review and meta-analysis. The Lancet Respiratory Medicine 4(2):116–28.

[103] Richardson TS, Robins JM (2013) Single World Intervention Graphs (SWIGs): A unification of the counterfactual and graphical approaches to causality. Center for the Statistics and the Social Sciences, University of Washington Series. Working Paper 128(30).

[104] Hernán MA (2015) Counterpoint: Epidemiology to guide decision-making: moving away from practice-free research. Am J Epidemiol 26:kwv215.

Supplementary Material

The approach advocated by the FACE applied to an investigation of urinary tract infections and low birthweight

Suppose we are interested in the potential causal relationship between urinary tract infections (UTIs) in the 2nd and 3rd trimesters of pregnancy and the subsequent low birthweight (LBW) of newborn babies, and that existing epidemiological evidence is suggestive of a clinically meaningful causal effect but not definitively so. Thus, a large prospective cohort study is designed, with laboratory assessment of the exposure in the second and third trimester of pregnancy.

In this appendix, we will outline how some of the key concepts from the FACE might be applied to the analysis of this cohort study and the interpretation of the results. We are not aiming to cover all possible concerns here; of course in any given real example there will always be specific additional features and complexities to be considered. Rather, we aim to highlight the thinking and the assumptions as they relate to a particular simple example, and above all to outline how the linking of the target causal estimand and the data can be achieved via these assumptions in a few simple mathematical steps. We recommend that those wanting to read about more realistic examples, with numerous additional complexities, turn to the many substantive papers cited in the main manuscript, *e.g.* [10–13, 43].

As with any epidemiological investigation, we would start (at the point of designing the study, more than when starting the analysis) with establishing what exactly is the question of interest. This involves asking:

A. What is the population of interest? All pregnant women living in the particular region/country to be studied? Perhaps non-singleton births should be excluded? And so on.

B. What exactly is the exposure of interest? Are all subtypes of UTIs to be included? What about those experiencing more than one UTI during pregnancy? What about the timing of acquiring the UTI(s)? *etc.*

C. How will the outcome be classified in pre-term births? *etc.*

D. Do we simply want to quantify the magnitude of the effect of maternal UTIs on the prevalence of LBW in the population of interest? Or are we interested in particular subgroups, *e.g.* particular racial groups, or mothers with gestational diabetes? Are we interested in investigating possible synergistic effects of other infections? And so on.

Let us assume for simplicity that the population of interest is all full-term singleton live-born babies from a particular region in a particular period, and that the exposure is simplified to ever having experienced a UTI in the second and/or third trimester of pregnancy versus never in these two trimesters of pregnancy. Let X be the binary exposure (maternal UTI, ever, coded 1, or never, coded 0), Y the binary outcome (LBW newborn, yes or no) and $\mathbf{C} = C_1, \ldots, C_p$ a set of potential confounders, *e.g.* maternal age, comorbidities (such as diabetes), other maternal infections and socio-economic factors.

The following features would characterise a FACE approach to this example:

1. **Estimand**: Express the question (or questions) of interest mathematically, *i.e.* as an estimand (or as estimands).

Potential Outcomes

Since causal questions involve not just features of the data at hand, but also a notion of *how things would have been had something been different*, then expressing causal questions mathematically requires an extension to the traditional statistical language of expectations, variances, probabilities, odds, *etc.* The extension most commonly used in the FACE is that of potential outcomes.

For each infant in the population of interest, let $Y(0)$ denote the outcome that would have been observed for that infant had the mother, possibly counter to fact, not been exposed during pregnancy, and let $Y(1)$ denote the outcome that would have been observed had the mother been exposed.

Possible Estimands

A population-level causal effect is then expressed as a contrast between the distributions of $Y(0)$ and $Y(1)$. For example, a marginal causal risk difference is:

$$E\left\{Y(1) - Y(0)\right\}$$

where expectations are taken over the population of interest. This estimand is interpreted as the difference in prevelance of LBW if, hypothetically, all pregnant women would be exposed, versus if no pregnant woman were exposed.

If we were interested in effect modification by maternal age, then we could compare the following conditional risk difference:

$$E\left\{Y(1) - Y(0)|A = a\right\}$$

for different values of a, where A is the age of the mother.

Suppose a different type of maternal infection, MIB, together with UTI, were believed to have a possible synergistic effect on LBW, then we would redefine our potential outcomes to include MIB too. $Y(0,0)$, $Y(1,0)$, $Y(0,1)$ and $Y(1,1)$ could be used to denote, respectively, the potential value that Y would take were both infections to be absent, only UTI present, only MIB present, or both exposures present. A synergistic effect of the two exposures would be present if:

$$E\left\{Y(1,1) - Y(0,1)\right\} > E\left\{Y(1,0) - Y(0,0)\right\}$$

i.e. if the effect of UTI is more pronounced when combined with MIB.

The estimand can also be chosen so as to reflect an interest in the excess prevalence of LBW that is attributed to UTIs, *i.e.*

$$E\left\{Y - Y(0)\right\}. \tag{1}$$

The possibilities are endless, but the resulting considerations are broadly similar, and so, for simplicity, in what follows, we assume that it is this final estimand that is of interest, *i.e.* the difference between the actual prevalence of LBW and the hypothetical prevalence of LBW that would be seen if UTIs could be prevented for all pregnant woman in the relevant population of interest.

Specificity

As we discuss in the main manuscript, for this estimand to be well-understood, the potential

outcome involved in it should ideally be close to being well-specified. This is not too problematic with an exposure such as infection. The hypothetical world in which no pregnant woman is infected could be one in which a hypothetical preventative treatment exists, with no (good or bad) side-effects, and is given to all pregnant women. It may also be useful to add that this hypothetical treatment be free for all pregnant women, so that in the hypothetical world in which all pregnant women are vaccinated, there are no knock-on effects on, say, their diet, due to having re-allocated resources towards acquiring the treatment.

For our chosen estimand, sufficiently well-specified $Y(0)$ would suffice. However, had we chosen an alternative estimand, we would also need to consider the meaning of $Y(1)$, and this could perhaps be a little trickier in this setting, since there could be many different reasons for acquiring a UTI present in the observed data. For simplicity, suppose there are two reasons: (1) high glucose level (*e.g.* due to poorly-controlled or undiagnosed diabetes), (2) incomplete voiding of the bladder. It may be plausible (from subject-matter knowledge) that the effect of maternal UTI on infant birthweight is the same regardless of the reason for its contraction, in which case the intervention (contract UTI by some means) would be sufficiently well-specified. However, if this were not the case, for example, if the UTIs acquired due to reason (1) were more severe than those acquired due to reason (2), then (for reasons that will become apparent below) the hypothetical intervention that we should imagine as giving rise to $Y(1)$ is a compound intervention (see [41]) in which the infection is contracted for reason (1) with probability π and for reason (2) with probability $1-\pi$, where π is the proportion in the observed data who actually contracted the infection due to reason (1). To be as specific as possible, it would be good to try to infer π from the data. If this is not possible, then the ambiguity should be made explicit.

2. **Assumptions**: Depending on the analytic approach to be taken, different assumptions may be invoked. For example, instrumental variable analyses rely on a different set of assumptions than analyses based on confounder adjustment. Suppose, for example, that we will proceed by confounder adjustment, then the assumptions under which the causal estimand (1) above can be identified (*i.e.* re-written as functions of the distribution of the observed data) are:

   (a) The *consistency* assumption, which for this example states that:

   $$Y = Y(0) \text{ if } X = 0.$$

   In words, this says that for infants whose mothers were in reality not exposed, their outcome in reality is equal to what it would have been in the hypothetical world in which no mother was exposed. For this assumption to hold, the observed data need to be *relevant* for the hypothetical intervention we have in mind that gives rise to $Y(0)$. Or, in other words, the hypothetical intervention we have in mind should be non-invasive in the sense described in the main manuscript: the hypothetical intervention, when applied to prevent UTIs, should not change the outcome for pregnant women who would in reality not have been exposed, from the outcome that was in fact observed. Recall the caveats above that the hypothetical treatment should have no (good or bad) side effects and should be free. These caveats help to make the consistency assumption more plausible; they precisely concern the non-invasiveness of the hypothetical intervention.

(b) The *conditional exchangeability* assumption, which for this example states that:

$$Y(0) \perp\!\!\!\perp X | \mathbf{C}.$$

This states that, having adjusted for the observed set of potential confounders $\mathbf{C}$, the exposure and the potential outcome $Y(0)$ should be independent. In this example, $Y(0)$ denotes whether or not an infant would have been LBW had his/her mother, possibly counter to fact, not been exposed. It can therefore be viewed as a relevant summary of *all the other* determinants of LBW (observed or unobserved) apart from UTIs. If the exposure were to be associated with these other determinants, even after conditioning on all the measured confounders, this would mean that there is residual confounding, *i.e.* even after allowing for all the measured confounders, if UTIs are, say, more common amongst those with a higher risk of LBW even in the absence of UTIs, then adjusting for the measured confounders will not lead us to a causally-interpretable estimate of (1). Causal diagrams are often advocated as a tool to help us decide how plausible the conditional exchangeability assumption is, based on existing knowledge (or plausible conjectures) regarding the causal structure of X, Y and $\mathbf{C}$. The more complex the situation (*e.g.* when variables contained in $\mathbf{C}$ are not simply common causes of X and Y) the more useful the causal diagrams are.

(c) The *positivity* assumption. For the case of our estimand, the relevant version is that, if all confounders $\mathbf{C}$ are discrete:

$$\text{For all } \mathbf{c} \text{ such that } P\{\mathbf{C} = \mathbf{c}\} > 0,$$

$$P(X = 0 | \mathbf{C} = \mathbf{c}) > 0.$$

In words, this says that, for a sufficiently large sample size, there should be unexposed mothers observed at every observed value of the confounders. For continuous confounders, a similar condition can be expressed using densities.

3. **Identification**: Under the assumptions above, we can re-write our casual estimand of interest (1) in terms of aspects of the joint distribution of the observed data. It is useful to follow these mathematical steps, to appreciate why we need the assumptions above. First we re-write our estimand using iterated expectations as:

$$E\{Y - Y(0)\} = E(Y) - E[E\{Y(0) | \mathbf{C}\}].$$

Then, under the conditional exchangeability assumption, $Y(0)$ is independent of X given $\mathbf{C}$, which gives us the licence to insert $X = 0$ on the right-hand side of the conditioning line in the second term, which means that our estimand is re-written as:

$$E(Y) - E[E\{Y(0) | X = 0, \mathbf{C}\}].$$

Then, under the consistency assumption, we can replace $Y(0)$ by Y on the left-hand side of the conditioning line in the same term, giving us:[6]

$$E\{Y - Y(0)\} = E(Y) - E\{E(Y | X = 0, \mathbf{C})\}.$$

[6]Note that consistency, as we stated it, is stronger than what is required to achieve this step. It would be sufficient to have consistency in conditional expectation given $\mathbf{C}$, *i.e.* that $E\{Y(0) | X = 0, \mathbf{C}\} = E(Y | X = 0, \mathbf{C})$. This relaxation would be important for the mixture intervention discussed for $Y(1)$ above, since consistency would be violated for such an intervention, but the weaker $E\{Y(1) | X = 1, \mathbf{C}\} = E(Y | X = 1, \mathbf{C})$ would be satisfies provided the interventions contributing to the mixture were all non-invasive.

This is important, since—using our first two assumptions—we have been able to write our estimand of interest, which involves data that we *don't* have, *i.e.* data from a hypothetical world in which there are no UTIs for pregnant women, in terms of aspects of the distribution of the data that we *do* have.

If all confounders were discrete, our estimand (1) could be estimated from a sufficiently large dataset on X, Y and $\mathbf{C}$ as long as the positivity assumption also holds. Positivity would ensure that $E(Y|X=0,\mathbf{C}=\mathbf{c})$ could be non-parametrically estimated, for every observed value $\mathbf{c}$ of the confounders, as the observed prevalence of LBW among infants with unexposed mothers with level $\mathbf{c}$ of the confounders. In practice, parametric (or semiparametric) estimation is more typically done, although this can be made very flexible using machine learning [33].

4. **Statistical estimation and inference**: There are many estimation options, some traditional and some novel proposed by the FACE—this is where propensity scores, inverse probability weighting, doubly-robust methods, and doubly-robust methods that incorporate machine learning algorithms come in—but the simplest option is to estimate $E(Y|X=0,\mathbf{C})$ directly from a parametric regression model fitted to the observed data. In this example, we could regress Y on $\mathbf{C}$ among the uninfected mothers, *e.g.* using a logistic regression model (including interactions/non-linearities as the data suggest, being liberal in as much as the sample size allows, since prediction of the potential outcome is the goal here, rather than interpretability of the model coefficients). We would then use the estimated coefficients from this model to predict the probability of $Y=1$, under hypothetically no infection, based on $\mathbf{C}$, for each individual in the dataset (uninfected and infected). The average of these predictions would be our estimate of $E\{E(Y|X=0,\mathbf{C})\}$. Subtracting this from the empirical mean of Y would give us our estimate of our causal estimand of interest, the prevalence of LBW attributable to UTIs in this population. We would also obtain the SE, CI *etc.* using either analytic or empirical (*e.g.* bootstrap) methods.

5. **Interpretation, misinterpretation and sensitivity analysis**: The effect estimates (and associated measures of statistical uncertainty) are interpreted, but, in particular, plausible violations of the assumptions are discussed explicitly in order to investigate possible misinterpretations. More formally, the sensitivity of the inference to possible violations of the assumptions can be assessed.

The FACE is of course concerned with a far larger set of problems than is represented by this very simple example, but the principles above carry through to settings with repeated exposures, time-dependent confounding, instrumental variables, effect modification, interaction, mediation, time-to-event outcomes, semiparametric estimation based on propensity scores *etc.*

A comparison with the 'traditional' approach

Traditionally, epidemiologists would certainly devote attention to establishing what is the question of interest. They would not usually, however, formally convert this into an estimand of interest that can be expressed mathematically as we did in feature 1. above. In some settings, things may be obvious enough for skipping 1. not to matter. However, skipping 1. can lead to serious mistakes in the later listed features. For example, the concepts of effect modification and interaction may not be distinguished, estimands such as the effect of the exposure *in the exposed* may not be easy

to articulate, and differences between possible direct effects (*e.g.* controlled vs. natural) may be overlooked. More seriously, without a clear mathematical estimand, there is the risk of calling the result of certain calculations, *e.g.* product of coefficients in mediation analysis, two-stage least squares in IV estimation, an indirect effect or a causal effect, respectively, even when such an interpretation is not allowed. Since feature 1. would traditionally be skipped, feature 3. (a mathematical link, via assumptions, between the estimand and the data) could not be attempted, and therefore a formal statement of assumptions (feature 2.) is not usually made. Traditionally, the assumption in 2(b)—no unmeasured confounding—would be stated, albeit more vaguely, but the consistency and positivity assumptions are usually overlooked. Without a clear idea of the assumptions made, sensitivity analyses (feature 5.) cannot reliably be carried out.