Brit. J. Phil. Sci. **70** (2019), 901–926

The Confounding Question of Confounding Causes in Randomized Trials

Jonathan Fuller

ABSTRACT

It is sometimes thought that randomized study group allocation is uniquely proficient at producing comparison groups that are evenly balanced for all confounding causes. Philosophers have argued that in real randomized controlled trials this balance assumption typically fails. But is the balance assumption an important ideal? I run a thought experiment, the CONFOUND study, to answer this question. I then suggest a new account of causal inference in ideal and real comparative group studies that helps clarify the roles of confounding variables and randomization.

1 *Confounders and Causes*
2 *The Balance Assumption*
3 *The CONFOUND Study*
  3.1 *CONFOUND 1*
  3.2 *CONFOUND 2*
4 *Disjunction C and the Ideal Study*
  4.1 *The ultimate 'other cause': C*
  4.2 *The ideal comparative group study*
  4.3 *Required conditions for causal inference*
5 *Confounders as Causes, Confounders as Correlates*
6 *Summary*

1 Confounders and Causes

The reasoning behind classical controlled experiments is simple. The researcher sets up an experimental condition and a control condition so that they are as alike as possible in every causally relevant way except for one experimental factor. If there is a difference in effect, then logic compels us

doi:10.1093/bjps/axx015

Advance Access published on January 22, 2018

to accept that the experimental factor is a cause of the effect. John Stuart Mill called this inference scheme the 'method of difference'.

Regrettably, human studies are not so simple. Comparative group studies, in which researchers compare human populations, deviate considerably from Mill's ideal, as human populations are heterogeneous in terms of variables that might be relevant but the researchers cannot manipulate (for example, age). When these variables are not balanced among the study groups, statisticians, epidemiologists, and social scientists tend to call them 'confounding factors' or 'confounders'. One clever way of 'controlling for' confounders is by allocating participants to the study groups such that the confounders are balanced among the groups, similarly represented within each.

One *a posteriori* strategy for achieving balance in comparative group studies involves looking for factors we suspect to be relevant and ensuring that the groups are 'matched' for each of these factors. This strategy is harnessed in observational group studies, a mixed bag of study designs in which subjects are generally followed in the course of routine life.

The *a priori* strategy involves randomly allocating participants to the study groups. If the study's sample is large enough (so the rationale goes), we can rest assured that confounders, including the ones we do not suspect, will be distributed evenly across the groups. In other words, we run a randomized controlled trial (RCT). From the second half of the twentieth century onwards, RCTs became increasingly popular in the human sciences. The evidence-based medicine (EBM) movement began in the early 1990s, and the evidence-based policy movement in education, social planning, and other areas followed on its heels. Both movements recommend that, whenever possible, decision- and policy-makers should use RCTs instead of observational studies when assessing the effectiveness of an intervention.

A treatment or policy intervention is effective only if it causes some outcome. Researchers thus measure effectiveness through a causal inference. In an RCT, randomization is thought crucially important for the causal inference; one influential claim sometimes made on its behalf is that randomization controls for all of the confounding variables, including those that are known (suspected) as well as those that are unknown (unsuspected).

John Worrall ([2002]), drawing on work by Peter Urbach (for example, Urbach [1985]), considers various arguments for the superiority of the RCT and concludes that it is overrated within EBM. Notably, he rejects the claim that randomization controls for all confounding variables, even probabilistically speaking. Yet Worrall worries that epidemiologists (Worrall [2002]) and philosophers (Worrall [2007]) commit to the assumption that—at least in the ideal—RCTs control for all confounding variables. Philosophers often worry about confounders as other causes of the study outcome—what I will call 'confounding causes'—that could explain a difference in outcome between

study groups if they too are distributed differently between groups. If RCT causal inference demands balance in confounding causes yet RCTs cannot supply it (with a reasonable probability), confounding causes seem to present a confounding conundrum at once philosophical and scientific.

In this article, my aim is to probe the importance of confounders in RCT causal inference. Doing so will throw light on the role of randomization, which is often at issue in debates about whether randomized studies are epistemically superior to observational studies. I call the assumption that all confounding causes are balanced among the study groups the 'balance assumption'. Some argue that RCTs, our gold standard studies, are usually unable to balance all confounders and thus the balance assumption fails (first commitment). Yet, intuitively, the balance assumption is an important ideal for comparative group study causal inference (second commitment). It thus seems that either our gold standard studies typically fall short of the ideal, or one of the previous two commitments is mistaken.

I will accept the first commitment in Section 2, and deny the second commitment in Section 3—the balance assumption is the wrong logical ideal. In Section 4, I will propose an alternate ideal, along with the required conditions for causal inference, based on a new account of causal inference. The ideal ought to guide the design of comparative group studies, while the required conditions ought to guide the interpretation of group study results. Finally, in Section 5, I will distinguish two concepts of 'confounder'. I will argue that confounders are primarily important not as causes but as correlates of the ultimate 'other cause', which I call '*C*'. The role of randomization is not to balance confounding causes, but to prevent systematic imbalances in these confounding correlates at baseline.

2 The Balance Assumption

There are three parts to the balance assumption in need of clarification: what a confounding cause is, what it means for a confounding cause to be balanced in a study, and what it means for all confounding causes to be balanced.

The concept of epidemiological confounding has undergone several historical revolutions (Morabia [2011]), and is a highly confused concept even its modern form (Greenland and Robins [1986]; Pearl [2009]). The confusion may be partly due to a clustering of multiple related yet distinct concepts under the term 'confounder'. One sense of confounder or confounding factor is an alternate cause of the study outcome (not the study exposure) that researchers must control for in order to avoid a faulty causal inference. This 'direct causal' concept of confounder is widely used in the philosophical literature on RCT causal inference (Papineau [1994]; Worrall [2002]; Howson and Urbach [2006]; Cartwright [2010]; Howick [2011]; La Caze [2013]). As Worrall ([2002],

pp. S321–2) argues, 'The effects of the factor whose effect is being investigated must be "shielded" from other possible confounding factors [...] There is, however, clearly an indefinite number of unknown factors that might play a causal role'. David Papineau ([1994], p. 439) cautions us to wonder whether 'some other confounding cause is responsible' for a measured association in a study. Meanwhile, Jeremy Howick ([2011], p. 35) describes conditions that a factor must satisfy in order to count as a confounding factor, including that 'the factor potentially affects the outcome'.[1] Finally, Colin Howson and Peter Urbach ([2006], p. 184) define 'prognostic factors' (a term often used interchangeably with confounding factors) as 'those respects that are causally relevant to the progress of the medical condition under study'.

An inventory of common confounding factors makes for a heterogeneous list. For instance, Howick ([2011]) mentions exercise, age, social class, health, and placebo effects as potential confounders. Exercise is believed to play a preventive causal role in mechanisms that produce cardiovascular outcomes like heart attack and stroke, and for this reason might be associated with these outcomes in a study. Yet it is not clear that the next two factors on the list are causes. Age is certainly a stock example of a confounding variable that randomization is supposed to disarm. Yet age, as the number of years elapsed since a participant was born, may only be associated with health outcomes because both age and health outcomes change over time, and these changes have a characteristic direction (for example, health outcomes tend to worsen). Similarly, it is arguable whether social class is a genuine cause, or if it is merely associated with causes like income and education. Variables like age, social class, postal code, and appreciation for classical music might all be associated (positively or negatively) with outcomes like heart attack or stroke, but it is not obvious that they causally affect those outcomes. They may fail Howick's own causal criterion for confounders. Thus there are variables that plausibly defy the direct causal concept of confounder yet are considered paradigmatic confounding factors nonetheless. I will distinguish confounders that are causes of the study outcome ('confounding causes') from those that are not. The balance assumption I will examine here refers exclusively to confounding causes.

The kinds of causes considered as potential confounders by philosophers and by scientists (diseases, genes, lifestyle factors, environmental exposures, demographic characteristics) are what epidemiologists Kenneth Rothman and Sander Greenland ([2005]) call 'component causes'. According to Rothman and Greenland, component causes are individual factors that interact within

[1] Howick's ([2011], p. 35) second condition for a confounder is an orthodox one: 'The factor is unequally distributed between experimental and control groups'. If this condition is not established or if it fails, we can call the factor a 'potential confounder'.

'complete causal mechanisms'. Exercise interacts with dietary and metabolic factors to produce healthy outcomes; genes and environmental exposures interact to produce diseases; and diseases and a lack of treatment interact to produce disease outcomes. Epidemiologists commonly represent complete causal mechanisms using 'causal pies', with component causes represented by slices within the pies. For now, I will understand confounding causes as component causes of the outcome (other than the study exposure) that are imbalanced among the study groups. In Section 4, I will fill in this sketch by exploring the logical relationship between confounding causes and study outcomes. Then in Section 5, I will turn to a distinct concept of confounder and explain its relevance for group study causal inference.

Worrall ([2002]) presumes that the notion of balance used by the methodologists he cites is statistical: a factor is imbalanced when its distribution is highly skewed and balanced otherwise. In other words, a variable is balanced when its average value or relative frequency is not significantly different between groups. What does it mean for all confounding causes to be balanced? For Worrall's methodologists, it means that each confounder (suspected or unsuspected) is balanced. Therefore, I will understand the balance assumption as maintaining that each potential confounding cause is distributed similarly among the study groups. A balanced distribution of each confounder is appealing because at first glance it is the kind of ideal comparability that warrants a causal inference in a comparative group study with positive findings. Although Worrall attributes to his sources the idea that randomization achieves balance in all variables, some confounders (for instance, placebo effects) are principally controlled through other common RCT design features (such as blinding).

Worrall ([2002]) reconstructs various arguments made by methodologists and philosophers as claiming that balance in all potential confounders probably (rather than certainly) obtains in an RCT.[2,3] After all, even if there is only one confounding cause and the process determining its distribution in the

[2] Howick ([2011], p. 50; my emphasis) also quotes Bradford Hill's *Principles of Medical Statistics* ([1991]) as claiming: 'We can equalise only for such features as we can measure or otherwise observe, but we also need unbiased allocation for *all* other features, some of which we may not even know exist. Only randomisation can give us that'. More recently, Edward Cox and colleagues ([2014], p. 2350; my emphasis) state that: 'Randomization *ensures* reasonable similarity of the test and control groups and protects against various imbalances and biases that could lead to erroneous conclusions'.

[3] In response to Worrall ([2002], [2007]), statistician Stephen Senn ([2013]) objects that his fellow statisticians are well aware that many potential confounders will be imbalanced in a randomized trial, and that the conventional analysis of trials assumes as much (see also La Caze *et al.* [2012]). He further argues that Worrall's concern about baseline imbalances is irrelevant because 'It is not necessary for the groups to be balanced' (p. 1442); to believe otherwise is to subscribe to a myth that Senn assumes 'no medical statisticians believe' (p. 1439). Even if Senn is right, because the balance assumption seems to function as an important ideal in philosophical accounts of RCT inference as well as in actual clinical research (as we will see), it is worthy of examination.

study is random, there is still a small chance that the distribution of that one cause is significantly skewed.

Just how probable is it that the balance assumption will obtain in any given RCT? Worrall ([2002], p. S324) argues that it is potentially unlikely: 'given that there are indefinitely many possible confounding factors, then it would seem to follow that the probability that there is some factor on which the two groups are unbalanced (when remember randomly constructed) might for all anyone knows be high'. He argues that it is a 'quantificational fallacy' to infer that the probability of balance in indefinitely many confounders is high from a high probability of balance in any one particular confounder. I accept Worrall's point that if there are an indefinite or unknown number of potential confounders, then the probability that all potential confounders are balanced is indefinite or unknown—and not necessarily high.

But perhaps we can be a bit more definite. Again, since Worrall describes the balancing for which randomization in particular is responsible, let us concentrate on the distribution of confounding causes at baseline. We can quantify the probability that all causes are balanced at baseline, p('all'). Assuming that the relevant causes are statistically independent of one another, p('all') = (p('one'))^n, where p('one') is the probability that one cause is balanced, and n is the number of unique causes. Let us also permit a weak degree of balance; say, a range of similarity in the distribution of the cause between groups that we would expect 95 times out of 100 when we randomize 'in the long run', so that p('one') = 0.95. Then p('all') = ('0.95')^n. If $n = 14$, then p('all') = 0.49. Thus, if there are fourteen or more unique confounding causes in an RCT, one or more will probably be imbalanced at baseline, even if the probability of balance is high for any one given cause. Fourteen is probably an underestimation of the number of statistically independent causes in a randomized trial. For instance, dozens of genes contribute to the endpoints in which health researchers and social scientists are interested, and most genes are inherited independently of one another. The upshot is that there is good reason to doubt that the balance assumption is true for even one of our best RCTs.

Worrall ([2007]) cites philosophers, including Nancy Cartwright ([1989]), who endorse the importance of controlling for all confounding causes in an RCT, cashed out in terms of probabilistic independence between group allocation and each cause. According to Cartwright ([1989], p. 64), in an 'ideal RCT', by definition the 'assignment of individuals to either the treatment or the control group should be statistically independent of all other causally relevant features that an individual has or will come to have'. Might there be a link between (i) probabilistic independence between group assignment

and each cause, and (ii) a balanced frequency distribution for each cause? Worrall ([2007], p. 472) conjectures:

> [...] if we were to take the study population and divide it again and again by some randomizing device into control and experimental groups and keep a cumulative total of the relative outcomes in the two groups, then we would expect that in the indefinite long run, the innumerable other possible causal factors would balance out [among study groups].

Worrall is pointing out that in a long-run RCT, the balance assumption might be satisfied. Along these lines, Papineau ([1994], p. 447)—whose account of RCT inference Worrall ([2007]) also discusses—claims that randomization ensures that all other causes of the outcome are probabilistically independent of the treatment, which will 'show up, not just in this sample, but in the long-run frequencies as the randomized experiment is done time and again'. But as Worrall argues, whatever may be true in a long-run or ideal RCT is not necessarily true in a real RCT. Thus, the probabilistic accounts that Worrall surveys provide us with no further reason to believe that the balance assumption will obtain in reality.[4] However, the question of whether ideal RCT causal inference should rely on the balance assumption remains. This question will occupy us in the next section.

3 The CONFOUND Study

So far we have seen that balancing all causes in an RCT is a lofty ideal. But why bother with these confounded confounding causes in the first place? Cartwright ([2011], p. 751) makes explicit one powerful intuition favouring this strategy: 'The underlying supposition is that differences in probabilities require a causal explanation; if the distribution of causes in the two groups is the same but for T yet the probability of O differs between them, the only possible explanation is that T causes O'. This supposition is based on the intuition that a difference in effect implies a difference in cause. To harness this inferential machinery, one runs a Mill's method of difference study. The method of difference is the usual paradigm for classical controlled experiments, but Cartwright ([2011], p. 751) suggests that observational studies and RCTs involve the same logic.[5]

[4] Nor should the accounts of Cartwright and Papineau be understood as describing the distributions of confounding causes that typically obtain in finite, real-world RCTs. Rather, they describe ideal sufficient conditions for causal inference. Cartwright ([2010], p. 64), citing (Worrall [2002]), concedes that 'it is of course not clear how closely any real RCT approximates the ideal'. In Section 4, I will examine the sufficient conditions that Cartwright proposes.

[5] J. S. Mill would not agree; he denied that the method of difference can be applied to group comparisons (Mill [1882]; Morabia [2013]).

For instance, in a case-control study the investigators compare a positive case of the outcome with a control case in which the outcome is absent. The choice of control is not made arbitrarily; the investigators select a control that is similar to the case in its causally relevant background circumstances. The investigators can then look for a potential cause of the outcome that was present in the positive case but absent in the control case. In a comparative group study, a balanced distribution of other causes seems to do the work that similarity in background causal circumstances does in a case-control study.

Of course, complete identity in all background causes, which the method of difference demands, is unlikely. Nonetheless, as Mackie ([1965]) argues, the method of difference is a logical ideal towards which scientists strive in their controlled experiments. Analogously, balance in each and every confounding cause, however improbable, seems to function as an ideal for our group comparisons, and our confidence in the soundness of our causal inference increases as the comparability of our groups increases. Embracing this idea, Howick ([2011]) recognizes that clinical trials are typically not sufficiently large to rule out all baseline confounders. But in response to Worrall ([2002], [2007]), he argues that this fact does not undermine the advantages of RCTs over observational studies: the former allow us to rule out a greater number of confounders than the latter, and it is on this basis that RCTs should be judged superior. Even though our randomized studies do not achieve the ideal, if they are closer to it than our non-randomized studies, then perhaps they are better after all.

I turn to now the question of whether we should in fact hold onto this ideal of balance in all confounding causes. In particular, is the balance assumption ever enough for sound causal inference in comparative group studies, and is it ever needed for sound causal inference? The first part of the question asks if balance in all confounding causes is sufficient for the conclusion that the exposure caused or prevented the outcome in a study showing a difference in outcome between groups. The second part asks if balance in all confounding causes is necessary for the causal conclusion in a study showing a difference in outcome. The following thought experiment will suffice to answer both questions. In the tradition of referring to clinical studies using a handy acronym, I will call this one the CONFOUND (CONceptual and epistemic FOUNDations of causal inference) study. The comparisons I am about to describe could be controlled trials of an intervention, but they could just as easily be observational group studies examining any kind of exposure, harmful or beneficial. The CONFOUND study actually includes two group comparisons: CONFOUND 1 and CONFOUND 2.

3.1 CONFOUND 1

CONFOUND 1 will examine a hypothesis that any comparative group study is designed to test: the exposure (X) caused the outcome (Y). There are, of course, other causes of Y (confounding causes) for which the investigators must control. For simplicity's sake, we will restrict the number of relevant confounding causes to two: C_1 and C_2. Also to make matters simple, X, Y, C_1, and C_2 are all dichotomous variables—that is, each variable is either present or absent in an individual participant. Each variable is measured as a frequency in the overall study group.[6]

Finally, I will make two deterministic assumptions. The first assumption is that causes act deterministically, that the set of causes present for an individual fully determines whether or not that individual gets the outcome. This simplification will allow us to rule out the possibility that any difference in the frequency of Y between groups is due solely to chancy causation. The assumption that causes determine their effects is traditionally called determinism, and can be summarized by the slogan, 'same (complete) cause, same effect'. But we can more precisely call it 'forward determinism', to distinguish it from a distinct deterministic assumption that I will also assume. The second assumption—call it 'reverse determinism'—is that whether or not an individual gets the outcome fully determines whether or not there was a complete cause of the outcome. Reverse determinism adheres to the slogan, 'some effect, some (complete) cause'. It discounts the possibility that any difference in the frequency of Y between groups is due solely to Y's spontaneously popping into existence, uncaused. Together, forward determinism and reverse determinism imply that any difference in the frequency of the outcome between groups is proof of some relevant causal difference. Despite these two simplifying assumptions, the lessons learned in this section will apply just as well to situations in which we do not assume determinism and compare probabilities instead of frequencies (as we will see in Section 4.2).

The CONFOUND 1 investigators measure the frequencies of Y, C_1, and C_2 in a group exposed to X, as well as in an unexposed group. Their results are presented in Table 1. As is typical, the investigators have incomplete background knowledge. In fact, all they know is that Y is the effect of one or more of X, C_1, and C_2, which exhaust the variables that are plausibly causally relevant. Table 1 is similar in many respects to a Mill's method of difference table, but while a Mill table typically has a '+' or '–' representing the presence or absence of a factor for an individual, Table 1 includes a number representing the frequency of a factor in a study group.

[6] In epidemiology, the absolute risk measures the relative outcome frequency, or the proportion of individuals with the outcome. In medicine, risks are often interpreted probabilistically (Fuller and Flores [2015]).

Table 1. Comparative group study in which all confounding causes are balanced. Numbers are frequencies; groups are equal in size.

	Y	X	C_1	C_2
Exposed	0.5	1.0	0.5	0.5
Unexposed	0	0	0.5	0.5

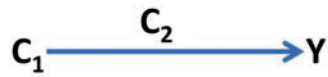

Figure 1. Mechanism producing Y in CONFOUND 1.

Despite their ignorance of the relevant causal mechanisms, the investigators see from Table 1 that (i) X is positively correlated with Y: an increased frequency of X is accompanied by an increased frequency of Y. They also see that (ii) all confounding causes of Y are balanced in the study; each confounding cause is (perfectly) uncorrelated with exposure X. They suppose that (iii) if all causes of Y are balanced except X and X is positively correlated with Y, then X must have caused Y. The investigators also happen to be disciples of Mill, so they reason using a kind of method of difference inference scheme: from (i), (ii), and (iii), they conclude that X caused Y in the study.

At this point in our thought experiment, we will allow ourselves to be omniscient and find out what really happened at the individual level. In each participant, Y represents the presence of a clinically important protein biomarker. The presence of Y is fully determined by the conjunction of C_1 and C_2. The confounders C_1 and C_2 represent two other proteins, each coded by different genes. C_1 is a precursor for Y, while C_2 is the enzyme that catalyses the conversion of C_1 to Y. The pathway is represented in Figure 1. X plays no part in this mechanism, which is the only mechanism that produces Y. Thus, X does not cause Y. How then can we explain the study results?

The key is that to say C_1 and C_2 are balanced is not to say all that much of use. Neither C_1 nor C_2 will cause Y without the other. Rather than the distribution of C_1 and the distribution of C_2 provided by Table 1, we need to know the distributions of $C_1\&C_2$, $C_1\&\neg C_2$, $\neg C_1\&C_2$, and $\neg C_1\&\neg C_2$. The frequencies for C_1 and for C_2 in Table 1 are consistent with a range of possible frequencies for these four conjunctions. It turns out that the actual frequencies are those reported in Table 2. This table reveals that in the exposed group, 50% of individuals were positive for both C_1 and C_2 ($C_1\&C_2$), which explains the 50% frequency of Y in that group because C_1 and C_2 are jointly sufficient for Y. However, in the unexposed group, the 50% of individuals who were

Table 2. Supplementary data for Table 1. Numbers are frequencies; groups are equal in size.

	Y	X	$C_1\&C_2$	$C_1\&\neg C_2$	$\neg C_1\&C_2$	$\neg C_1\&\neg C_2$
Exposed	0.5	1.0	0.5	0	0	0.5
Unexposed	0	0	0	0.5	0.5	0

positive for C_1 (column '$C_1\&\neg C_2$') were not the same 50% of individuals who were positive for C_2 (column '$\neg C_1\&C_2$'). Because neither C_1 nor C_2 will cause Y without the other, no one in the unexposed group was positive for Y. Exposure X plays no role in this causal story. Yet the researchers thought that it must because all confounding causes were balanced and there was a difference in outcome between the groups!

Even those who doubt the likelihood of balancing all confounding causes in a randomized trial sometimes accept the sufficiency of this condition. For instance, Howson and Urbach ([2006], p. 197) suggest that a guarantee that the comparison groups are balanced for each prognostic factor 'has at least the virtue that if it were true, then the conditions for an eliminative induction would be met, so that whatever differences arose between the groups in the clinical trial could be infallibly attributed to the trial treatment'. What the hypothetical CONFOUND 1 study shows is that the balance assumption is not enough for sound causal inference. In a comparative group study with positive results, the finding that all confounding causes are balanced is not sufficient for inferring that X caused Y, as demonstrated by the folly of our black box researchers.

3.2 CONFOUND 2

Although the balance assumption is not sufficient, it might perhaps be necessary. It might be indispensable for sound causal inference, and thus a crucial consideration for trialists. To investigate this possibility, let us commission a second thought study, CONFOUND 2. This time, let us hypothesize that another exposure (X^*) caused outcome Y. The researchers in this study are in much the same situation as before: they are told that C_1 and C_2 are the only plausible confounding causes, and that they can make deterministic assumptions, but they are given no other information. They measure the frequencies of X^*, C_1, C_2, and Y in a study population (Table 3). Once more, the researchers observe that (i) X^* is positively correlated with Y. They still maintain that (iii) if all causes of Y are balanced except X^* and X^* is positively correlated with Y, then X^* must have caused Y. However, it is now not that case that (ii) all confounding causes of Y are balanced in the study: C_1 is extremely imbalanced. Thus,

Table 3. Comparative group study in which not all confounding causes are balanced. Numbers are frequencies; groups are equal in size.

	Y	X^*	C_1	C_2
Exposed	0.5	1.0	0.75	0.5
Unexposed	0.25	0	0.25	0.5

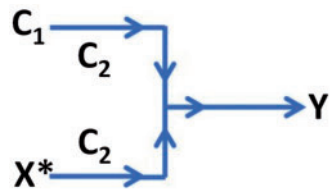

Figure 2. Mechanism producing Y in CONFOUND 2.

Table 4. Supplementary data for Table 3. Numbers are frequencies; groups are equal in size

	Y	X^*	$C_1\&C_2$	$C_1\&\neg C_2$	$\neg C_1\&C_2$	$\neg C_1\&\neg C_2$
Exposed	0.5	1.0	0.25	0.5	0.25	0
Unexposed	0.25	0	0.25	0	0.25	0.5

they do not conclude that X^* caused Y in the study. They were reasonable not to do so, given their epistemic disadvantage. We, on the other hand, are able to take a look inside the black box and find out what really happened.

The human pathway involving the four variables is represented in Figure 2. As before, the C_2 enzyme always catalyses the conversion of C_1 to Y. This time, X^* is very similar to C_1, so C_2 will also catalyse the conversion of X^* to Y. Table 4 explains the results of Table 3 in light of this mechanism. We know from the biological mechanism that what is most salient is the frequency of individuals who are positive for $C_1\&C_2$. This frequency is the same in both groups, $C_1\&C_2$ is perfectly balanced. In the unexposed group, the 25% of individuals who were positive for $C_1\&C_2$ fully account for the 25% frequency of Y. In the exposed group, the 25% of individuals who were positive for $C_1\&C_2$ accounts for half of all individuals who were positive for Y. The remaining 25% of Y-positive participants in the exposed group must have gotten the outcome via some other causal pathway. Since C_1 and C_2 on their own cannot cause Y, the only other difference—X^*—must be involved. Indeed, X^*

caused Y in the other 25% of participants with C_2 in the exposed group, who are found in the column '$\neg C_1 \& C_2$'.

This reasoning reveals that the balance assumption is not necessary for sound causal inference. Even if the assumption fails, we might still be able to confidently infer that an exposure caused a difference in outcome between groups, if only we had the right information. In summary then, the balance assumption is not enough and not needed; in a study with positive results, the truth of the balance assumption is neither necessary nor sufficient for a sound causal inference.

Unfortunately, the balance assumption may function as an ideal not only in philosophical accounts of RCT inference but also in real-world research. Yusuf *et al.* ([1990], p. 77) note: 'By using as an analogy experiments conducted in a test tube or in animals, it is often argued that all extraneous and confounding variables can and should be controlled'. Britton *et al.* ([1999], p. 117) suggest: 'In the classical laboratory experiment, the effect of the variable of interest is isolated by controlling the values of other relevant variables [. . .] there may be a residual feeling that an RCT is a form of laboratory experiment'. Both articles argue that the quest to control all confounders is misguided, and perhaps arises through inappropriate analogy with classical (Mill's method of difference) experiments.

A balanced distribution of each confounding cause is the wrong logical ideal for making inferences in comparative group studies. Consequently, we should not appraise the epistemic worth of a group study—whether a randomized trial or non-randomized study—according to how closely it approaches this ideal. However, we are not done with confounders and causes just yet. As I will show in the next section, the idea of balancing 'other causes' is not too far off the mark.

4 Disjunction *C* and the Ideal Study

Returning to the confounding conundrum I posed at the outset, I accepted that in an RCT it is unlikely that all confounding causes are balanced among the study groups; the balance assumption probably fails. However, we need not worry yet because we have also seen that the balance assumption is not the ideal condition we might have thought it was for making causal inferences in comparative group studies: it is neither necessary nor sufficient.

In Section 4.1, I will introduce two new concepts: complex causes and *C*. These concepts are useful for understanding comparative group study causal inference, but also causation in epidemiology and the social sciences more generally. I will then propose a new account of causal inference in comparative

group studies, distinguishing the ideal conditions (Section 4.2) from the required conditions (Section 4.3) for causal inference.

4.1 The ultimate other cause: *C*

It is well documented among philosophers and scientists that what we ordinarily call causes are components of more complete causal mechanisms. In the world of epidemiology, Rothman and Greenland ([2005]) distinguish 'component causes' from the complete mechanisms, which they call 'sufficient causes'. Mackie ([1965], [1980]) developed a nomenclature to describe the logical relations among causes and their effects. In Mackie's language, the conjunction A&B&C is minimally sufficient for D just when all components (A, B, C) are jointly sufficient for D, but no subset of the components (neither A, nor B, nor C, nor A&B, nor A&C, nor B&C) is sufficient for D. Similarly, Rothman and Greenland ([2005], p. S144) define a sufficient cause of disease as 'a set of minimal conditions and events that inevitably produce disease; "minimal" implies that all of the conditions or events are necessary to that occurrence'. A conjunct in a minimally sufficient condition can be a negated term, representing an 'interfering factor' or 'counteracting cause'. What I will call a complex cause or sufficient cause has similar properties; it can be described by a conjunction ($C_1\&C_2\&\ldots$) of non-redundant single (negated or unnegated) terms, where $C_1\&C_2\&\ldots$ is minimally sufficient for causing Y.

Although Mackie's minimally sufficient conditions fully determine the effect, our definition of a complex or sufficient cause should not be understood as precluding the possibility of chancy causes, those that act indeterministically. A complex cause can be minimally sufficient for causing Y even if the chance of Y occurring given the complex cause is less than 100% (think of a radioactive atom causing the stochastic emission of an alpha particle). As Papineau ([1985]) suggests, sufficiency can be expressed here in terms of the cause determining the chance of the effect rather than the effect itself.[7]

Component causes are parts of complex causes; they are represented by the conjuncts. We can now define confounding causes as component causes within complex causes that exclude the exposure under consideration. On their own, confounding causes are insufficient for causing the outcome in a study. Neither precursor proteins, nor enzymes, nor exercise, nor placebos are sufficient causes. This fact explains why confounding causes are only derivatively important: they are important insofar as they constitute complex causes, but it

[7] I also do not intend to rule out the possibility that Y is a quantitative variable. Cartwright ([2012]) adapts Mackie's definition to account for quantitative or multivalued effect variables: instead of being sufficient for the effect, a complex might be sufficient for producing a contribution to the value of the effect.

is controlling for complex causes that is of ultimate importance in causal inference.[8]

Both Mill and Mackie were aware that there exists a 'plurality of causes' (Mackie [1980], p. 307) for any phenomenon (Rothman and Greenland [2005] use the term 'multicausality'). There are many sufficient causes—each uniquely constituted—of the effect, Y. In the CONFOUND study, we considered only one complex cause that excluded the exposure (C_1&C_2). In any real study, there will be a great many more. At first glance, the plurality of complex causes might pose a problem for the comparability of our groups. If something like the balance assumption—but for complex causes rather than confounding causes—is desired, then we run into the same initial problem that plagued the balance assumption, namely, given the number of unique complex causes in a group study, it might be improbable that each unique complex cause is distributed equally among the groups. Fortunately, what is needed for sound causal inference is something much weaker than a balancing of each unique complex cause.

To help understand why, let us define a disjunction, C, where C includes all unique complex causes of Y except any that involve exposure X as a conjunct: $C = (C_{1,1}\&C_{1,2}\&\ldots) \vee \ldots \vee (C_{n,1}\&C_{n,2}\&\ldots C_{n,m})$.[9] The conjuncts in this general formula (for example, $C_{1,1}$) are potential confounding causes, while the disjuncts (for example, $C_{1,1}\&C_{1,2}\&\ldots$) are complex causes (note that $C_{n,1}$ may or may not be identical to $C_{1,1}$). Since C is a disjunction of complex causes, it is satisfied whenever an individual has any sufficient cause of Y (except those containing X, by stipulation). C is the ultimate 'other cause'. Assuming determinism, what matters for sound causal inference is not the distribution of each unique complex cause, because any sufficient cause will cause Y. What matters is the distribution of participants satisfying C, which abstracts away the particulars. In the next section, we will explore what matters for causal inference if we do not assume determinism.

Think back to CONFOUND 2 and Table 4. Sufficient cause C_1&C_2 was perfectly balanced between the study groups, which allowed us to causally attribute the difference in outcome to the exposure (X*). Since C_1&C_2 was the only sufficient cause of the outcome that did not include the exposure, we can fill in the general formula for C using only one conjunction: $C = C_1\&C_2$ for this exposure and this outcome. If instead there were two sufficient causes, it

[8] Judea Pearl ([2009], p. 195) examines traditional no-confounding criteria and similarly challenges the sufficiency and necessity of the assumption that all potentially relevant variables are unassociated with the exposure. Pearl proposes as a solution the notion of a 'non-trivial sufficient set' that bares similarities to my concept of a sufficient or complex cause.

[9] Or $C = \mathrm{V}_i\Lambda_j C_{ij}$. I am assuming that Y is a dichotomous variable. For quantitative effect variables there is an analogue of the dichotomous C term, which we can call 'quantitative C'. Quantitative C is a function of confounding causes, while the quantitative Y variable is a function of quantitative C (and of X when X causes Y).

Table 5. Ideal conditions for a comparative group study causal inference. Numbers and variables are distributions. Assume all complex causes have an equal effect on Y, $y_X > y_{\neg X}$.

	Y	X	C
Exposed	y_X	1.0	z
Unexposed	$y_{\neg X}$	0	z

would not matter if each complex cause was greatly imbalanced so long as the disjunction of both sufficient causes (C) was not.

4.2 The ideal comparative group study

In the CONFOUND study, causal conclusions follow deductively from (i) the positive association between exposure and outcome, (ii) the premise that C is distributed equally between groups, and (iii) our two deterministic assumptions, forward determinism and reverse determinism.[10] To see how, consider Table 5, representing ideal conditions for a comparative group study causal inference. We need only consider three factors: Y, X, and C. Forward determinism (same (sufficient) cause, same effect) ensures that every instance of a complex cause produces the outcome, and thus the frequency of Y in the unexposed group can be no less than the frequency of C. Reverse determinism (some effect, some (sufficient) cause) guarantees that every instance of the outcome is produced by a complex cause, so that the frequency of Y in the unexposed group can be no more than the frequency of C. Together, these two deterministic assumptions imply that the frequency of C in the unexposed group is equal to the frequency of Y: $z = y_{\neg X}$. Because C is distributed perfectly evenly between the groups, the frequency of C in the exposed group is also equal to $y_{\neg X}$.

The final special feature of the ideal study represented by Table 5 is that the frequency of the outcome is greater in the exposed group compared to the unexposed group ($y_X > y_{\neg X}$). In other words, the study shows a 'positive result'.[11] Recall that we already worked out that the frequency of C in the exposed group is equal to $y_{\neg X}$ in our ideal study. Therefore, in the exposed group the frequency of Y ($=y_X$) is greater than the frequency of C ($=y_{\neg X}$).

[10] Whenever X causes Y via a causal mechanism that includes downstream confounding causes, C will occur (that is, X causes Y by causing C). X can thereby cause imbalances in C that would not bias our causal inference were we to conclude that X caused Y. To avoid false negative inferences, we can stipulate that it is the conjunction of (i) C and (ii) the absence of a complex cause involving X that is distributed equally between study groups. From now on, I will speak only of C's distribution in ideal and real studies and will assume conjunct (ii).

[11] The finding that $y_X < y_{\neg X}$ could also be regarded as a positive result, suggesting that X prevents Y. I do not have space to discuss the logic of prevention here.

There must be some individuals who got outcome Y but lacked a complex cause in C. According to the assumption of reverse determinism, these outcomes must have had some cause. If not C, each of these outcomes must have been caused by one of the complex causes excluded from C: complex causes involving X. Thus, exposure X is a cause of the outcome. Furthermore, we can causally attribute the difference in outcome between study groups to the exposure.

I have relied on a deterministic ideal only to make vivid the deductive validity of group study causal inference. This reasoning is equally valid if we allow for indeterministic causation and measure the probabilities of Y and C. Instead of considering the distributions in Table 5 as frequency distributions, we can consider them as probability distributions. Instead of forward determinism, assume that the probability of Y is completely determined by the set of causes present for an individual (the probability of Y is zero in the absence of a complex cause and greater than zero in the presence of a complex cause). Then any difference in the probability of Y between groups must be due to some relevant causal difference. If C is distributed equally (C is probabilistically independent of the exposure), then the relevant causal difference must involve the exposure, and we can causally attribute the difference in the probability of the outcome to the exposure. In measuring the probability of C rather than the probability of each complex cause in C, I am assuming (for simplicity) that each complex cause determines the same probability of Y. I will have more to say about this assumption shortly.

Cartwright ([2010]) also articulates sufficient conditions for causal inference to undergird her ideal RCT. She starts by defining subpopulations that are each homogeneous with respect to the combination of causally relevant factors (confounding causes) that are present. In the CONFOUND study, there were four unique subpopulations: $C_1\&C_2$, $C_1\&\neg C_2$, $\neg C_1\&C_2$, and $\neg C_1\&\neg C_2$. According to Cartwright ([2010], p. 64), 'In an ideal RCT each K_i [subpopulation] will appear in both [study] wings with the same probability'. Then if there is a higher probability of the outcome in the treatment group ($p(Y|X) > p(Y|\neg X)$), these conditions entail that the treatment causes the outcome in at least one of the subpopulations. Cartwright's conditions for causal inference are sufficient by the lights of my account, as some of the subpopulations will contain complex causes in C and these subpopulations will each be equally distributed between the groups. However, Cartwright's ideal RCT is more demanding than my ideal study (Table 5) for two reasons. First, it requires that unique subpopulations not containing complex causes in C are each distributed equally, which is not needed. Similarly, in Cartwright's ideal RCT each unique subpopulation containing a complex cause in C is distributed equally, which is also stronger than necessary, especially if we assume that each complex cause fixes the same probability of Y.

If we allow that two unique complex causes might fix two unique probabilities of the outcome (a reasonable allowance), then we should instead demand that C's contribution to the probability of Y is the same for both groups, or that C's contribution is balanced. C's contribution to the probability of Y is the average probability of Y among C subpopulations, multiplied by the total probability of C subpopulations. We can think of it as the force that C exerts on Y. If C's contribution to the probability of Y is the same for both groups, yet the total probability of Y—the net force on Y—is greater in the exposed group ($p(Y|X) > p(Y|\neg X)$), then X is causally responsible for this difference in probability.[12,13]

The condition that C's contribution be balanced is the key criterion of an ideal study, requiring the most metaphysically modest assumption among those that we have been working with: the probability of Y is completely determined by the set of causes present. Whenever unique complex causes fix unique probabilities of the outcome, C's contribution can 'balance out' in several ways, just as there are multiple ways to balance a scale using weights of different masses. If instead all unique complex causes fix the same probability of the outcome, C's contribution is balanced whenever the probability of C is the same in all study groups; and if complex causes fully determine the outcome, C's contribution is balanced whenever the frequency of C is the same in all study groups—whenever C is balanced.

To summarize, a balanced distribution of all confounding causes is a confounded (confused) logical ideal. On the other hand, a balanced contribution

[12] The proof of this principle requires some work. First, we must define a 'C subpopulation' as a homogeneous subpopulation with at least one complex cause in C and without any complex causes that are not contained in C. Then we can define 'C's contribution to the probability of Y ($C_{P(Y)}$)' as the (weighted) average probability of Y among C subpopulations, multiplied by the total probability of C subpopulations. If there are two unique complex causes ($complex_1$ and $complex_2$):

$$C_{p(Y)} = p(Y|C) \times p(C)$$

$$C_{p(Y)} = [p(Y|complex_1)p(complex_1|C) + p(Y|complex_2)p(complex_2|C)] \times [p(complex_1) + p(complex_2)].$$

The total probability of Y in a study group is the sum of C's contribution to Y and $\neg C$'s contribution to Y, or:

$$p(Y) = p(Y|C)p(C) + p(Y|\neg C)p(\neg C)$$

$$p(Y) = C_{p(Y)} + p(Y|\neg C)p(\neg C).$$

The $p(Y|\neg C) = 0$ in the unexposed group because the $\neg C$ partition of the unexposed group contains no complex causes (none in C, none involving X), and—by the assumption of reverse determinism—Y can only ever occur when it is caused. If $C_{p(Y)}$ is the same in both groups, yet the total probability of Y is greater in the exposed group, then $p(Y|\neg C) > 0$ in the exposed group. This is only possible if the $\neg C$ partition of the exposed group contains some complex causes involving X, in which case X causes Y in one or more subpopulations of the exposed group.

[13] A similar principle applies when Y is a quantitative effect variable and a quantitative C variable represents the contribution to Y of causes that do not interact with X (for a linear model, see Cartwright [2012]). In this case, in the ideal study the average effect of quantitative C on quantitative Y is the same in all study groups.

of C among study groups is an unconfounded ideal, free from confusion and free from epidemiological confounding. In a comparative group study showing an association between exposure and outcome, the premise that C's contribution is balanced between groups is sufficient for our causal conclusions (so long as we make the necessary metaphysical assumptions).[14]

The causal conclusions to which I am referring (the exposure is a cause of the outcome, the difference in outcome in the study is causally attributable to the exposure) are somewhat modest. They tell us nothing about whether the exposure will cause the outcome in a different population, such as a relevant target population for an intervention. Nor do they quantify the exposure's effect size in the overall study population (the aggregate of all study groups), let alone in any other population. These inferences require further assumptions that I cannot develop here.[15]

4.3 Required conditions for causal inference

So far I have argued that perfect balance in C's contribution to Y is sufficient for sound causal inference in a positive group study. I have not shown that perfect balance is required for the causal inference, and in fact it is not. To conclude that X caused Y, all we require is that C's contribution is less imbalanced than Y's distribution. If we make deterministic assumptions, then whenever C is less imbalanced than Y, the Cs cannot account for all of the difference in Y, which leaves only X to account for some of the difference. If we instead assume that each complex cause fixes the same probability of Y, we can conclude that X caused Y when the ratio of C's probability in the exposed group to its probability in the unexposed group is less than the ratio of Y's probabilities ($\mathrm{p}(C|X)/\mathrm{p}(C|\neg X) < \mathrm{p}(Y|X)/\mathrm{p}(Y|\neg X)$). Whenever C's probability is less imbalanced than Y's probability, C cannot fully account for the difference in Y's probability, and X must be partly causally responsible. To see this, we can insert any value we like for Y and C in Table 5 (easing the requirement that C is evenly distributed), and suppose any values we want for

[14] If C's contribution is balanced and there is no difference in outcome between groups, it is not necessarily true that the exposure did not cause the outcome. It could be the case—however unlikely—that the exposure caused the outcome in some participants and prevented the outcome in just as many participants.

[15] Roughly, the effect size quantifies the difference in outcome that the exposure makes in the population. In order to accurately predict the effect size in a population of interest, the distribution of the exposure's 'support factors' in the exposed group must be properly representative of their distribution in the population of interest. The support factors are those causes that interact with the exposure to cause the outcome. In CONFOUND 2, enzyme C_2 was a support factor for X*. Because the distribution of $\neg C_1 \& C_2$ in the exposed group was representative of its distribution in the overall study population, the difference in outcome between groups (0.25) accurately predicts the effect size in the overall study population.

the probability of Y given C and the probability of Y given $\neg C$ (I will leave this exercise to the reader).

Although it may be unrealistic and unnecessary for sound causal inference, a perfectly balanced contribution of C serves an important function as a regulative ideal for the design of a comparative group study. The various techniques and tricks used in comparative studies—randomization, double-blinding, stratification, matching—are an attempt to bring the distribution of C closer to the ideal, so that we can feel more confident that any difference in outcome is due to the exposure. However, we should not despair that real studies typically fall short of the ideal; all that is required in the interpretation of the study's result is that C's contribution is less imbalanced than the outcome.[16] Much of the philosophical literature on RCT causal inference has focused on inferences in ideal studies; but in interpreting the results of real studies, we should not let the ideal be the enemy of the sufficient.

In summary, rather than a balance in confounding causes, group study causal inference depends upon a balanced contribution of C. We should have more confidence in our causal inference whenever we are more confident that C's contribution is less imbalanced than the outcome. Any of the complex causes in C are sufficient for causing the study outcome. Meanwhile, confounding causes are only conjuncts of complex causes in C—on their own, they are not sufficient for the outcome. Are philosophers of science then wrong about the importance of confounders in group studies? We will now see that they are not wrong; they have simply exaggerated the importance of confounders as causes. Recognizing the importance of confounders as correlates of C helps clarify the role of randomization in group studies.

5 Confounders as Causes, Confounders as Correlates

Though we have relieved the tension between the implausibility of balancing each confounding cause and the apparent need to do so in a group study, there is something left unresolved: why statisticians are so worried about confounding variables. So far, we have examined the role of confounders as causes of the study outcome, a role that philosophers often assume in their accounts of RCT causal inference. In this section, we will analyse a concept of confounder

[16] My account of comparative group study causal inference is most related to a regularity theory of causation of the kind proposed by Mackie ([1980]), given the similarity of my sufficient causes to Mackie's minimally sufficient conditions, as well as my reliance on regularity assumptions like reverse determinism. In comparison, the classic 'potential outcomes' approach to causal inference (Rubin [1974]; Greenland and Robbins [1986]) is closely related to counterfactual theories of causation, while causal Bayes nets approaches (Spirtes *et al.* [1993]; Pearl [2009]) are tied to interventionist theories.

distinct from the 'direct causal concept': confounders as correlates. More precisely, confounders are important as correlates of *C*.

Earlier, I noted that philosophers typically cast confounders as causes, but that some of our stock examples of confounders may not fit the bill. For example, it is not obvious that age and social class are causal variables. They are, however, associated with outcomes of interest—for instance, older patients are more likely to have atherosclerosis as a result of the accumulation of plaque in their arteries over time, and are thus more likely to have a heart attack or stroke.

In contrast to the philosophers, methodologists in epidemiology and the social sciences have historically worried about confounders not simply because they are (sometimes) causes but crucially because they are correlates of the outcome. Rothman and Greenland ([1998], p. 120) state: 'In general, a confounder must be associated with both the exposure under study and the disease under study to be confounding'. Meanwhile, Guyatt *et al.* ([2008], p. 777) define a confounder as: 'A factor that is associated with the outcome of interest and is differentially distributed in patients exposed and unexposed to the [exposure] of interest'. This textbook also discusses the importance of balancing prognostic factors (potential confounders): 'If prognostic factors—either those we know about or those we do not know about—prove unbalanced between a trial's treatment and *control groups*, the study's outcome will be biased' ([2008], p. 70). In other words, confounders can lead us to falsely conclude that the exposure caused or prevented the outcome when in fact it did not.

Associational definitions of 'confounder' are often unclear or incomplete. Where must a factor be 'associated with the outcome of interest' before it qualifies as a confounder? Presumably, in the comparative group study if its imbalance is to bias the study's results. However, an association between an imbalanced variable and the study outcome is not enough. If the treatment causes an excess of the outcome in the treatment group, any imbalanced variable will be automatically associated with the outcome; but the result is not thereby biased. What Guyett *et al.* are truly worried about are variables that are associated with instances of the outcome not caused by the treatment. But to say this is just to say that they are worried about variables associated with instances of the outcome caused by complex causes in *C*. In fact, we often suspect that a certain variable might be a confounder in a comparative study like a trial once we have observed its association with the outcome in a prognostic study—that is, a study searching for correlations between variables like age and outcomes like heart attack in a population that is not exposed to the trial treatment. If we assume that the heart attacks in this untreated population all had some cause, those causes must be found in *C*. Thus, if age is associated with the outcome in this untreated population, it is associated

with *C*. If the association between age and *C* also holds in the trial and age is imbalanced between trial groups, we should worry that *C*'s contribution is imbalanced. In short, confounders like age are important not as causes of the outcome but as correlates of *C*.

I have referred to the concept of a confounding cause as a 'direct causal' concept because, on this interpretation, a confounder is causally relevant to the outcome in a direct sense: it causes the outcome. We can consider the distinct concept of 'confounder as a correlate of *C*' to be an 'associational-causal' concept of confounder: confounding variables are factors associated with complex causes of the outcome in *C*. To avoid ambiguity, from this point on I will reserve the term 'confounding factor' for the direct causal concept, and label the associational-causal concept with the term 'prognostic factor' or 'covariate'. (Of course, many variables are both a prognostic factor and a confounding cause.) Though I have argued that the importance of balancing confounding causes in a group study has been exaggerated, I do not wish to deny a role for the direct causal concept of confounding causes in our causal reasoning. It may sometimes be background knowledge—our understanding of the confounding causes and how they mutually interact—that reveals a potential imbalance in *C*'s contribution rather than statistical correlations.

How might a covariate, a correlate of *C*, wind up imbalanced between the study groups? One mechanism involves chance: in a randomized study, randomization can throw up a highly unequal distribution. There are also non-random or systematic ways in which prognostic factors come to be imbalanced. In a non-randomized trial or in an observational group study, if the investigators or care providers can select which patients will receive the treatment of interest and which patients will not, for conscious or unconscious reasons they might treat with the new drug those patients that are on average younger or healthier. Then age or health will be associated with the drug, and as a result the drug might be associated with *C*. In this case, it is said that the study suffered from selection bias.

Properly executed randomization prevents selection bias in a controlled trial by preventing selection. Study investigators and participants are prevented from deciding group assignment; instead, group allocation is determined by a random process. Thus, randomization militates against systematic imbalances in *C* by barring systematic imbalances in prognostic factors at baseline. Of course, the possibility of a large baseline imbalance due to chance remains and, as we saw in Section 2, a chance imbalance is probable if we think that the number of relevant variables is great enough. Fortunately, just as we should not worry about balancing all confounding causes, we should not be concerned about the dismal prospects of balancing all of *C*'s correlates, of which there will be many. Like confounding causes, prognostic factors—as correlates of *C*—are only derivatively important. They serve an

epistemic function, alerting us to imbalances in C. It is the distribution of C itself that is of ultimate concern in causal inference, and C is only one variable. Once we have controlled for all systematic sources of imbalance in a trial (through randomization and subsequent methods), the chance that C's contribution is distributed more-or-less evenly between groups is relatively high, so long as the trial population is relatively large.

In an observational group study, the investigators typically do not dictate which patients receive the exposure because usually the data are collected in routine practice. Although they cannot prevent provider selection or patient self-selection, observational studies are not defenceless against selection bias. Investigators can match the study groups for similar distributions of prognostic variables, stratify the study groups according to prognostic variables and analyse results within strata, or make statistical adjustments to the data based on observed imbalances in prognostic variables. However, as the common refrain goes, these methods can only control for variables that are observed and that we suspect are relevant. The methods leave open the possibility of a hidden association between exposure and C that is not predicted by the observed, suspected prognostic variables. If we get a positive study result, we might then lack confidence that C's contribution is less unequally distributed than the outcome, or we might have difficulty assessing how confident we can be. How worried we should be about unforeseen selection bias should depend on how complete our knowledge is about the relevant variables, which will vary by circumstance.

A virtue of my account of comparative group study causal inference is that it proposes a common logic for randomized and non-randomized studies. In comparison, a causal Bayes nets account of causal inference in clinical research that includes an 'ideal intervention' on the exposure variable (for example, Steel [2011]) is more straightforwardly applicable to a randomized study compared with an observational group study, because it requires that patient characteristics (endogenous variables) do not influence exposure. A unified account of comparative group study causal inference allows us to compare randomized studies with non-randomized studies directly (in general and in particular cases). We can compare our confidence in studies that are quite distinct by holding them up to a singular ideal. In the design of any comparative group study, we strive towards the ideal of a properly balanced distribution of C among study groups. In interpreting the data from any study, our causal inference is secure if C's contribution is less imbalanced than the outcome. The diverse methods of RCTs and observational studies—randomizing, matching—can be seen as promoting a balanced distribution of C or as correcting for imbalances in C (via prognostic factors).

When we examine the ideal and required conditions for causal inference, we see that the matter of whether or not a study was randomized is not a premise in either inference. We can make a causal inference in any comparative study

when the study groups are sufficiently comparable, regardless of how we go about generating and assessing comparability. In understanding just what kind of comparability is needed, confounding causes are merely pixels in a more complex causal picture.

6 Summary

It is unlikely that a randomized trial—even if well designed and conducted—will achieve a balanced distribution for each confounding variable. Yet balance in all confounders, construed as confounding causes of the study outcome, is sometimes held up as a logical ideal, our confidence in our causal inference increasing as our comparative group study approaches the ideal. It turns out that the balance assumption is a false idol. Instead, we must worry about the distribution of sufficient or complex causes of the outcome that do not involve the study exposure (those included in *C*). In the ideal comparative study, *C*'s contribution is balanced among study groups. In any real study, if *C*'s contribution is less imbalanced than the outcome, then we can conclude that the exposure caused the outcome. Confounders are primarily important not as causes of the outcome but as correlates of *C* (prognostic factors or covariates) that may alert us to an imbalance in *C*'s contribution. Randomization prevents systematic imbalances in prognostic factors at baseline, which may improve the comparability of our study groups. But on the account of comparative group study causal inference presented here, it is the comparability of the study groups with respect to *C* that is of direct relevance.

Acknowledgements

I am grateful to Nancy Cartwright, Luis Flores, Bennett Holman, Ayelet Kuper, David Papineau, Jacob Stegenga, David Teira, Paul Thompson, Ross Upshur, Sarah Wieten, the anonymous reviewers, and audiences at the Philosophy of Science Association Biennial Meeting (2014), King's College London, University of Genoa, and the University of Toronto for inspiration, feedback, and discussion. I am thankful for support from the Canadian Institutes of Health Research and the W. Garfield Weston Foundation.

Faculty of Medicine
University of Toronto
Ontario, Canada
and
African Centre for Epistemology and
Philosophy of Science
University of Johannesburg
South Africa
jonathan.fuller@mail.utoronto.ca

References

Britton, A., McKee, M., Black, N., McPherson, K., Sanderson, C. and Bain, C. [1999]: 'Threats to Applicability of Randomised Trials: Exclusions and Selective Participation', *Journal of Health Services Research and Policy*, **4**, pp. 112–21.

Cartwright, N. [1989]: *Nature's Capacities and Their Measurement*, Oxford: Oxford University Press.

Cartwright, N. [2010]: 'What Are Randomised Controlled Trials Good For?', *Philosophical Studies*, **147**, pp. 59–70.

Cartwright, N. [2011]: 'Predicting "It Will Work for Us": (Way) beyond Statistics', in P. M. Illari and J. Williamson (*eds*), *Causality in the Sciences*, Oxford: Oxford University Press.

Cartwright, N. [2012]: 'Will This Policy Work for You? Predicting Effectiveness Better: How Philosophy Helps', *Philosophy of Science*, **79**, pp. 973–89.

Cox, E., Borio, L. and Temple, R. [2014]: 'Evaluating Ebola Therapies: The Case for RCTs', *New England Journal of Medicine*, **371**, pp. 2350–1.

Fuller, J. and Flores, L. J. [2015]: 'The Risk GP Model: The Standard Model of Prediction in Medicine', *Studies in History and Philosophy of Biological and Biomedical Sciences*, **54**, pp. 49–61.

Greenland, S. and Robins, J. M. [1986]: 'Identifiability, Exchangeability, and Epidemiological Confounding', *International Journal of Epidemiology*, **15**, pp. 413–19.

Guyatt, G., Rennie, D., Meade, M. O. and Cook, D. J. [2008]: *Users' Guides to the Medical Literature: Essentials of Evidence-Based Clinical Practice*, New York: McGraw-Hill Medical.

Hill, A. B. and Hill, I. D. [1991]: *Bradford Hill's Principles of Medical Statistics*, London: Edward Arnold.

Howick, J. [2011]: *The Philosophy of Evidence-Based Medicine*, Oxford: Wiley-Blackwell.

Howson, C. and Urbach, P. [2006]: *Scientific Reasoning: The Bayesian Approach*, Chicago, IL: Open Court Publishing.

La Caze, A. [2013]: 'Why Randomized Interventional Studies', *Journal of Medicine and Philosophy*, **38**, pp. 352–68.

La Caze, A., Djulbegovic, B. and Senn, S. [2012]: 'What Does Randomisation Achieve?', *Evidence-Based Medicine*, **17**, pp. 1–2.

Mackie, J. L. [1965]: 'Causes and Conditions', *American Philosophical Quarterly*, **2**, pp. 245–64.

Mackie, J. L. [1980]: *The Cement of the Universe: A Study of Causation*, Oxford: Oxford University Press.

Mill, J. S. [1882]: *A System of Logic, Ratiocinative and Inductive*, New York: Harper and Brothers.

Morabia, A. [2011]: 'History of the Modern Epidemiological Concept of Confounding', *Journal of Epidemiology and Community Health*, **65**, pp. 297–300.

Morabia, A. [2013]: 'Hume, Mill, Hill, and the *sui generis* Epidemiologic Approach to Causal Inference', *American Journal of Epidemiology*, **178**, pp. 1526–32.

Papineau, D. [1985]: 'Probabilities and Causes', *Journal of Philosophy*, **82**, pp. 57–74.
Papineau, D. [1994]: 'The Virtues of Randomization', *British Journal for the Philosophy of Science*, **45**, pp. 437–50.
Pearl, J. [2009]. *Causality: Models, Reasoning, and Inference*, New York: Cambridge University Press.
Rothman, K. J. and Greenland, S. [1998]: *Modern Epidemiology*, Philadelphia, PA: Lippincott-Raven.
Rothman, K. J. and Greenland, S. [2005]: 'Causation and Causal Inference in Epidemiology', *American Journal of Public Health*, **95**, pp. S144–50.
Rubin, D. B. [1974]: 'Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies', *Journal of Educational Psychology*, **56**, pp. 688–701.
Senn, S. [2013]: 'Seven Myths of Randomisation in Clinical Trials', *Statistics in Medicine*, **32**, pp. 1439–50.
Spirtes, P., Glymour, C. and Scheines, R. [1993]: *Causation, Prediction, and Search*, New York: Springer.
Steel, D. [2011]: 'Causal Inference and Medical Experiments', in D. M. Gabbay, P. Thagard and J. Woods (*eds*), *Philosophy of Medicine*, Amsterdam: Elsevier, pp. 159–85.
Urbach, P. M. [1985]: 'Randomisation and the Design of Experiments', *Philosophy of Science*, **52**, pp. 256–73.
Worrall, J. [2002]: 'What Evidence in Evidence-Based Medicine', *Philosophy of Science*, **69**, pp. S316–30.
Worrall, J. [2007]: 'Why There's No Cause to Randomize', *British Journal for the Philosophy of Science*, **58**, pp. 451–88.
Yusuf, S., Held, P., Teo, K. K. and Toretsky, E. R. [1990]: 'Selection of Patients for Randomized Controlled Trials: Implications of Wide or Narrow Eligibility Criteria', *Statistics in Medicine*, **9**, pp. 73–83.