Causal Inference in Biomedical Research

Tudor M. Baetu

(forthcoming in Biology & Philosophy)

Abstract

Causation can be inferred by two distinct patterns of reasoning, each requiring a distinct experimental design. Common, non-statistical causal inference is associated with controlled experiments in basic biomedical research. Statistical inference is associated with Randomized Controlled Trials in clinical research. The main difference between the two patterns of inference hinges on the satisfaction of a comparability requirement, which is in turn dictated by the nature of the objects of study, namely homogeneous vs. heterogeneous populations of biological systems. This distinction entails that the objection according to which randomized experiments fail to provide better evidence for causation because randomization cannot guarantee comparability is mistaken. As far as the validity of the statistical inference is concerned, randomization is not required in order to ensure comparability, but rather to prevent systematic bias which may compromise the accuracy of the intervention.

1. A debate concerning the virtues of randomization

Clinical trials aim to determine whether a medical intervention is a causal difference maker in respect to a health-related outcome in a patient or, more often, a population of patients. Randomized Controlled Trials (RCTs) are widely regarded as the gold standard in clinical research, providing the strongest evidence for causal efficacy (The Cochrane Collaboration 2011). It is not surprising, therefore, that one of the most debated questions is how exactly and to what extent the

main feature that differentiates RCTs from other types of controlled experiments, namely the random allocation of subjects to the test and control arms of the experiment, contributes to the validity of causal inference.

A common answer is that randomization ensures comparability, that is, an even distribution of potential confounders among patients in the test and control groups (Cartwright 2010; Papineau 1994). Critics challenge this claim, pointing out that randomization can balance the effects of confounders only in the long run, by performing an infinite series of experiments in which patients are randomly allocated to test and control conditions (Howson and Urbach 2006; Lindley 1982; Urbach 1985; 1993; Worrall 2007b). It would seem therefore that critics too, assume that valid statistical inference in clinical research requires comparable groups balanced in respect to potential confounders and that the main purpose of random allocation is to achieve comparability. However, since randomization cannot guarantee comparability, they conclude that randomized studies are not epistemically superior to non-randomized ones.

But why is comparability so crucially important to statistical inference? In this paper, I argue that if comparability can be assumed, then causation can be validly inferred without any further reliance on statistical testing. A controlled experiment contrasting a test and a control condition suffices. Conversely, the purpose of statistical testing is to provide a basis for inferring causation when the possibility that incomparable groups are contrasted cannot be discarded. Thus, there are two distinct methods for inferring causation: a common, or non-statistical, causal inference associated with controlled experiments in basic biomedical research; and a statistical inference associated with RCTs in clinical research. The main difference between the two forms of causal inference hinges on the satisfaction of a comparability requirement, which–as I will argue in the paper–is ultimately dictated by the nature of the objects of study, namely homogeneous vs. heterogeneous populations of biological systems. Researchers opt for a method of inference or the other, implementing the corresponding experimental design, depending on whether they have reasons to believe that a population is homogeneous or not. This distinction has an immediate consequence for the debate concerning the virtues of randomization: the assumption that comparability is required for the validity of statistical inference is mistaken. According to the account defended in the paper, the role of randomization vis-à-vis the validity of causal inference is not to promote comparability, but rather to ensure the accuracy of the intervention by removing some forms of systematic error.

The paper is organized as follows: In Section 2, I explain what I take to be the two methods for inferring causation. In each case, I analyze the inferential logic and the associated experimental requirements, highlighting some key differences. In Section 3, I focus on the role random allocation plays in respect to comparability, the accuracy of the intervention, and the validity of statistical tests. I argue that random allocation has no bearing on the validity of statistical tests and, although there is a relationship between random allocation and comparability, this relationship is not essential to statistical inference. Then, I show how the 'controlled experiment' and 'statistical testing' elements fit together in the experimental design of an RCT, and argue that random allocation is primarily required in order to ensure the accuracy of the allocation intervention. Section 4 summarizes the main claims made in the paper.

2. Two methods for inferring causation

2.1 Causal inference when comparability can be assumed

The most common test for demonstrating causation in basic biomedical research is the controlled experiment. The test consists in contrasting outcomes in two conditions that differ in respect to a tested variable. In order to yield a verdict regarding the causal relevance of the tested variable vis-à-vis the outcome, the experiment must further satisfy the following desiderata:

(i) An intervention on the variable (i.e., an experiment) must be conducted. Manipulation–as opposed to mere observation of differences in the outcome between two conditions–is standardly required in order to establish the directionality of causation (i.e., demonstrate that the variable is causally relevant to the outcome rather than the other way around) and rule out the possibility that the changes in the variable and the outcome are correlated due to a common cause. If changes in the variable and the outcome are divergent effects of a common cause or if the former is an effect of the latter, then the manipulation of the variable is not expected to have an impact on the outcome. However, if there is a causal pathway linking the tested variable and the outcome as upstream cause to downstream effect, then interventions on the variable are expected to result in changes in the outcome.

(ii) The test and the control conditions must be comparable in all relevant respects except for the variable manipulated in the experiment. Failure to ensure comparability raises the possibility that some other difference between the two conditions (a confounder) is responsible

for the observed differences in outcome. Comparability demands: (ii.1) that possible confounder-variables take the same values in the test and control conditions at the onset of the experiment; and (ii.2) that the experiment is shielded from unequal external interferences other than the intervention targeting the test condition. In other words, comparability is meant to ensure that the test and control systems start in the same state and evolve in the same way except for whatever changes are brought about by the intervention and its downstream effects. An allowance is made for the possibility that the values of causally relevant variables (including confounders) may change during the experiment, due to external inferences or to the natural progression of the system, and that these changes may affect the measured outcome. However, whatever these changes are, their impact is expected to be essentially identical in the two arms of the experiment.

(iii) The intervention should be accurate, in the sense that it should target only the variable under investigation. Accuracy is required to demonstrate the causal relevance of the tested (independent) variable to the differences in outcome. If the accuracy of the intervention cannot be demonstrated, it is still possible to demonstrate that the intervention itself is causally relevant. However, the causal efficacy of the intervention may be attributed to the fact that the intervention targets some other variable in addition to or instead of the tested variable.

Causal inference follows the general pattern of reasoning outlined in Mill's method of difference (1843, Chapter VIII, § 2), namely that of a contrastive inference whereby those aspects of a situation known to be constant (i.e., controlled) between two situations are ruled out as possible explanations of differences in outcome. (i) is meant to rule out the possibility of a non-causal correlation between the tested variable and the observed differences in outcome, as well as the possibility of a reverse causation scenario. (ii) is meant to rule out explanations of differences in outcome other than the intervention. (iii) is meant to further rule out explanations involving variables other than the designated independent variable which might have been affected by the intervention. If conditions (i) (ii) and (iii) are satisfied, there is good evidence to conclude that the tested variable is the causal difference-maker responsible for the observed differences in outcome.

Although causal inference is sometimes framed in terms of a probabilistic theory of causation [e.g. (Cartwright 2010)], this is not the form in which it is encountered in the experimental practice of basic biomedical research. Attributing probabilities to measured outcomes is by no

means a trivial task and there are no methodological guidelines specifying how such attributions should be made. Moreover, the probabilistic glossing should not overshadow the fact that the overall logic of causal inference is one of systematic elimination of alternative explanations, a point which is repeatedly emphasized in the methodological literature.[1] Finally, it is perhaps worth pointing out that since the causal inference is contrastive in nature, the claim 'X causes Y' is shorthand for 'a change in variable X is causally relevant to a difference in outcome Y between two situations given the context of a comparable background of other possible determinants.' In particular, 'X causes Y' should not be taken to imply that an entity or event X produces or 'brings about' an entity/event Y, or that X or some change in a variable X is a sufficient cause of Y or a change in a variable Y.

A detailed justification of the intervention requirement (condition i) can be found in the philosophical literature on interventionist accounts of causation (Pearl 2000; Spirtes et al. 1993; Woodward 2003). Since it is not essential to my argument, I will not reiterate it here.

In the experimental practice of the life sciences, comparability (condition ii) has two components, one referring to the living systems under investigation, the other to the replication of the experimental background in the test and control conditions. Experimental practices deployed to ensure the latter include the standardization and operationalization of the techniques of measurement and intervention. A parallel experimental design in which test and control are simultaneously deployed side by side is commonly adopted to ensure that causal interferences external to the experimental setup have an equal impact on both conditions (condition ii.2).

Ensuring the comparability of biological systems at the onset of the experiment (condition ii.1) is far more challenging. In basic science, the preferred strategy is to systematically remove differences between biological systems in order to generate genetically and phenotypically homogeneous organism strains and cell-line clones, which are then maintained in standardized living conditions (Ankeny and Leonelli 2011; Baetu 2016; Clarke and Fujimura 1992; Müller-Wille 2007). Further precautions are taken in order to ensure that test and control biological systems are

[1] "Determining whether there is a causal relationship between variables, A and B, requires that the variables covary, the presence of one variable preceding the other (e.g., A → B), and ruling out the presence of a third variable, C, which might mitigate the influence of A on B" (Leighton 2010, 622). Confounders are conceptualized as rival (usually causal) explanations of the observed difference in outcome between test and control. Thus, if the test and control systems differ in terms of factors that can impact on the measured outcome, the causal inference is deemed inconclusive since the possibility that something other than the manipulated factor may explain the difference in outcomes cannot be ruled out (Chow 2010).

'synchronized' (e.g., cells are at the same stage of the cell cycle). Thus, while most research in basic science relies on comparisons between populations of biological systems, such as cells and organisms, these populations are highly homogeneous and are often assumed to consist of quasi-identical copies of the same biological system. In addition to contributing towards the satisfaction of the comparability requirement, homogeneity also allows researchers to extrapolate causal claims from populations to individuals and vice versa without incurring a substantial risk of error.

The accuracy of the intervention (condition iii) is ensured by subjecting a technique of intervention to a process of validation meant to demonstrate that the technique targets only the variable under investigation and no other variables that may have similar features or contribute to similar differences in outcome. Validity is in part demonstrated by including additional positive and negative controls in experiments. For example, it is common practice to perform placebo interventions (pipetting, mixing, centrifuging, incubating) in the control arm of the experiment. Such interventions are meant to ensure that the relevant difference maker is not some generic lab procedure, such as gently shaking the cells, but rather the investigated variable, say, a virus, which is added by gently shaking the cell suspension.

2.2 Statistical inference in the context of RCTs

Homogeneous populations of biological systems are, by and large, generated in the laboratory. Natural populations, on the other hand, are notoriously heterogeneous. For instance, the progression and severity of medical conditions vary among patients. Moreover, even when a population of patients displays identical symptoms, the underlying physiopathology may still vary from one patient to the next. As a result, a treatment may be successful in some patients, but have no effect or even adverse effects in other patients. This makes it extremely difficult to satisfy the comparability requirement demanded by the method of inference described in Section 2.1.[2]

The variability of biological outcomes in a population is standardly modelled and explained as the effect of a multitude of causal factors, genetic and environmental, unevenly distributed

[2] The validity of extrapolations is also compromised. Heterogeneous populations and individuals drawn from these populations are no longer interchangeable experimental surrogates. If the individuals in a population differ in respect to confounding causal factors, causal claims established by studying individuals may not be representative of causal relationships prevalent in the general population, while causal relationships shown to be predominant in a population may not apply to a particular individual drawn from that population. In practical terms, this means that a treatment working well for a group of tested patients may turn out to have little impact in the general population, while a treatment generally successful in a population may not be effective for a particular patient.

among individuals in a population (Fisher 1947). A measure of the variability of an outcome could be a proportion, such as the fatality ratio associated with an untreated medical condition documented over a long period of time in a large population of diagnosed patients. If the fatality ratio (say, within two years following diagnosis) is 100%, it can be inferred that none of the patients with the condition ever recover on their own. In this case, the causal 'background noise' of confounders–i.e., causes other than the tested treatment that may contribute to recovery–is negligible. A sample of any size taken from this population is expected to be characterized by the same 100% fatality outcome. Conversely, the seemingly miraculous recovery of a single patient cannot be dismissed as random variation, since, in this idealized scenario, there is no variation in the outcome. Thus, if an intervention is conducted, then any instance of recovery can reasonably be attributed to the intervention and not to some other cause.

In contrast, if the fatality ratio is 20%, then the 'background noise' of confounders is quite substantial, with 80% of the patients recovering in the absence of any treatment. The fatality ratios observed in samples taken from this population are bound to vary depending on which individuals happen to be picked in each sample. In turn, variability makes it impossible to draw a valid conclusion about the efficacy of a treatment intervention by simply comparing the fatality ratio in a sample of treated patients with the fatality ratio in the general population.

Statistical testing provides a means to calculate the probability that differences in outcome between a sample and the general population are due to sampling alone.[3] If the sampling method is not biased–for instance if a random sampling technique is successfully implemented–it is possible to calculate the frequency with which sample fatality values occur by repeatedly picking samples of the same size. This can be easily illustrated for samples of two.[4] If a patient is picked at random from the population, there is a 20% chance of picking a patient who dies and an 80% chance of picking a patient who recovers. In the long run, 64% of the samples would display a fatality ratio of 0% (the probability that both patients recover is 0.8×0.8), 32% a ratio of 50% (one patient dies and the other survives or vice versa, $2 \times 0.2 \times 0.8$), and 4% a100% ratio (the probability that both patients die, 0.2×0.2). Thus, we can reason that even if a completely inefficacious treatment is administered to a sample–an explanation known as the 'null hypothesis'–, there is a 64% probability of observing a 0% fatality ratio because the sample happens to consist of patients who

[3] The distinction between statistical (or chance) and causal explanations is discussed in (Witteveen forthcoming).
[4] The example is adapted from (Hill 1955, Ch. VIII).

would have recovered anyway. It is still possible that the treatment is in fact efficacious and cured two patients who would have not recovered on their own (the probability of picking such a sample being 4%). Unfortunately, we don't know which kind of sample was tested. The only thing we know is that it is highly likely that we picked a sample in which the patients recover in the absence of any treatment (64%). Since we have no grounds to dismiss the null hypothesis, we should abstain from drawing any conclusions about the efficacy of the treatment.

As sample size increases, the probability of observing a 0% fatality ratio in virtue of random sampling alone decreases radically, although it is always possible that a sample, even a very large one, happens to consist predominantly of patients that would have recovered on their own. For a sample of 20, the probability of picking at random only patients that would have recovered anyway drops to 1.15% (0.8^{20}). Assuming that a treatment was administered and that all 20 patients recovered, we are entitled to conclude that it is rather unlikely that we picked a highly unrepresentative sample consisting only of patients who would have recovered on their own. It is more likely that something else, perhaps the treatment, is responsible for the recovery of the patients.

It is important to emphasize that statistical inference does not guarantee the truth of the conclusion even if all the assumptions underlying the inference (e.g., random sampling) are satisfied. Statistical testing only provides a means of quantifying the risk of drawing an erroneous conclusion. In order to infer causation, we must first decide whether the risk is acceptable or not. The agreed upon methodological convention is that a null hypothesis with a probability higher than 5% constitutes an unacceptable risk. This kind of risk and the associated methodological convention are not present in the common causal inference described in Section 2.1: if conditions (i)-(iii) are met, we are entitled to infer without any further ado that the manipulated variable is the cause of the observed differences in outcome. Of course, it may, and probably does happen that despite our best efforts the test and control systems are incomparable or that the intervention is inaccurate, but this concerns the satisfaction of the assumptions underpinning valid inference, not the inference itself.

In clinical practice, population statistics are usually unavailable. It is also difficult to randomly sample a population spread over a large geographical area. A different, more practical approach must be thought out. One strategy is to start with a non-random sample gathered by a technique in which the probability of getting any particular sample from a population cannot be calculated–this is known as a convenience sample–, describe it in as much detail as possible and then

extrapolate findings to a conceptually defined population consisting of all patients, present and future, similar to the study participants.

A version of this strategy is adopted in RCTs. First, the treatment and its method of delivery, as well as the outcome of interest and the method of assessment are defined. On the same occasion, the desired characteristics of potential study participants are specified by a set of eligibility criteria. Inclusion criteria include the diagnosis of the medical condition targeted by the tested treatment. Exclusion criteria are necessary for ethical and legal reasons, and to eliminate known confounders such as advanced age and comorbidities, which may mask improvements of the targeted medical condition. Then, a convenience sample of patients satisfying these criteria is gathered. Steps are taken to avoid known mechanisms of sampling bias, such as volunteerism, and the characteristics of the participants and the circumstances of their enrollment are recorded in detail; this information is subsequently used to evaluate extrapolations to other populations and individuals. Enrolled patients are then randomly allocated to either a test group, which receives the treatment under investigation, or a control group, which receives no treatment, a placebo or the standard treatment. Finally, outcomes are measured according to the agreed method of assessment.

Since not all patients diagnosed with the condition of interest satisfy the eligibility criteria, a legitimate concern often voiced by critics is that the reference population is not identical to the general population of patients (Howson and Urbach 2006, 190). Thus, even if a treatment is shown to be efficacious in the test group, the external validity of the trial, that is, the effectiveness of the treatment in the general population under routine healthcare conditions, remains uncertain. Despite this shortcoming, it is important to point out that both basic science and clinical research work on the thus far fruitful methodological assumption that biological systems are modular, such that it is possible to treat dysfunctions independently of one another. Among other things, eligibility criteria are applied in order to demonstrate the efficacy of the treatment in respect to a particular dysfunction. Thus, explanatory trials–that is, trials on patients satisfying eligibility criteria conducted under carefully monitored conditions–provide crucial information necessary to sustain basic and clinical research (La Caze 2013). This said, it is also possible to conduct pragmatic trials–that is, trials on patients drawn from the general population under routine healthcare conditions (Godwin et al. 2003)–, which can both directly test effectiveness in the general population and indirectly assess the modularity assumption.

A second concern is that the extrapolation from samples to population is not based on statistical inference, but relies on an informal notion of similarity (Howson and Urbach 2006, 190-91; Worrall 2007a, 994-95). This is true, but it doesn't follow that similarity-based extrapolations are invalid. They are extremely common in science and in many cases successful (Baetu 2016; Steel 2007).

Moreover, statistical testing is still used to infer causation in a way that does not depend on the notion of similarity used to generalize causal claims. To see how this is done, it is useful to consider once again a simple example. Let us suppose that an RCT is conducted to test the efficacy of a treatment, with 10 patients allocated to a placebo (control) group and 10 patients to a treatment (test) group. The outcomes (death or recovery) for each patient are listed in Figure 1, panel A. Overall, a 10% (1 out of 10) fatality ratio is observed in the treatment group, as opposed to 20% (2 out of 10) in the placebo group (panel C). The fact that the fatality ratio is relatively low in the control condition means that most patients recover whether or not they receive a treatment. Could it be then that the observed differences in outcome between the test and control groups (i.e., a 10% reduction in fatality) is due to the fact that more patients who recover on their own happened to be allocated to the test group? To evaluate the risk of making an incorrect causal inference, one needs to establish how often a difference in fatality ratios of this magnitude could arise solely by randomly assigning the patients to the test and control conditions. A randomization test, such as Fisher's exact test of independence, assumes the null hypothesis that there is no association between allocation and outcome. In other words, any given individual would incur the same outcome no matter the group to which is allocated. If this hypothesis is correct, then the 'treatment' and 'placebo' labels are interchangeable. The test therefore consists in erasing the labels and repeatedly reallocating at random (rerandomizing) the 20 patients to two groups of 10, taking note of the difference in fatality between the groups and the frequency with which each difference-value occurs. For instance, in one round of rerandomization we obtain the outcomes listed in panel B and summarized in panel D. After 1000 rerandomizations, we observe that a difference in fatality of 10% between treatment and placebo groups can easily occur in virtue of random allocation alone, with a probability of 49.7% (panel E). Since the null hypothesis cannot be ruled out, this particular experiment is inconclusive. The incertitude can be reduced by increasing group size. For groups of 100, the probability of observing a difference in fatality of 10% or larger in virtue of random allocation alone drops to 3% (panel F); for groups of 1000, the probability is close to zero (panel

G). Thus, if a 10% difference in fatality is observed for large groups, we would be entitled to reject the chance explanation and conclude that the allocation to the treatment group makes a difference in respect to recovery.

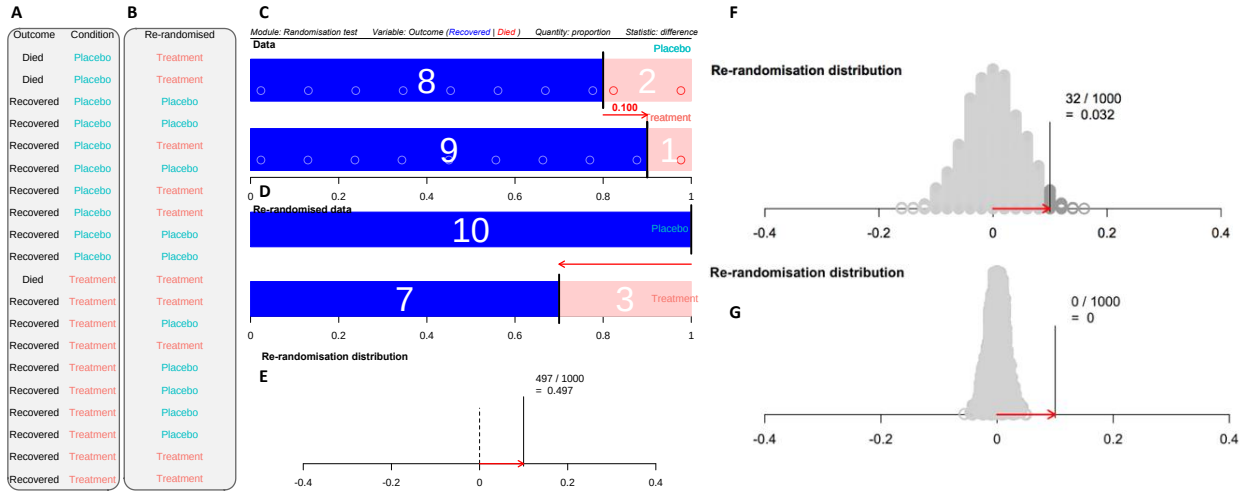

Figure 1. VIT (https://www.stat.auckland.ac.nz/~wild/VIT/) simulation of 1000 rerandomizations. One-tailed p-values for differences in fatality of 10% or more are indicated in panels E-G.

Just like the common, non-statistical causal inference discussed in Section 2.1, the overall logic of statistical inference is one of systematic ruling out of rival explanations (Fisher 1947; Hill 1952; 1955). In both cases, non-causal correlations are ruled out by conducting an experimental intervention (condition i) and spurious attributions of causal efficacy are ruled out by taking steps in order to ensure the accuracy of the intervention (condition iii; more on this in Section 3.4). The main difference concerns the comparability requirement (condition ii), which cannot be satisfied in studies involving individuals drawn from natural populations. In order to overcome this difficulty, one strategy for eliminating alternative explanations–namely, relying on comparability in order to rule out explanations involving differences other than interventions on the tested variable–is replaced by an inference based on the results of a statistical test assessing the probability that differences between groups are generated in virtue of random allocation alone. In turn, this has important consequences for experimental design and data analysis: statistical analysis is required, sufficiently large samples or groups are needed in order to reduce the risk of error to acceptable levels, and, as I will show in Section 3.4, randomization too is required in order to avoid spurious causal attribution.

3. The roles of comparability and randomization in statistical inference

3.1 The debate

Randomization, understood as the experimental practice of random allocation, is often described as an effective method for ensuring a high degree of comparability. For instance, we are told that:

> "Random assignment of subjects to the treatment or control wings […] is in aid of ensuring that other possible reasons for dependencies and independencies between cause and effect under test will be distributed identically in the treatment and control wings." (Cartwright 2010, 63)

Similar recommendations abound in the clinical literature. Take, for instance, this quote from a comprehensive series of review articles on the evaluation of trial results:

> "To ensure 'fair' comparison between the treatments, the different study groups must be truly comparable. This can be achieved by standardization of, for example, the time(s) of intake of the study medication and the methods used to measure clinical parameters, but most important for comparability is randomization of the participants." (Kabisch et al. 2011, 664)

Ultimately, such claims can be traced to Fisher's (1947, 19) treatment of the "procedure of randomisation" which he describes as the method "by which the validity of the test of significance may be guaranteed against corruption by the causes of disturbance which have not been eliminated."

Critics, on the other hand, claim that randomization is useless (and, moreover, should be banned for ethical reasons which I will not discuss here). One of their key arguments is that randomization can balance the effects of confounding factors only in the long run, by performing many experiments in which the patients are randomly allocated to test and control conditions (Howson and Urbach 2006; Lindley 1982; Urbach 1985; 1993). Thus,

> "the best that might be argued is that 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 and the limiting cumulative relative outcome would reflect the true efficacy of the treatment." (Worrall 2007b, 472)

Note that both arguments for and against randomization presuppose, first, that there is a logical link between randomization and comparability; and, second, that comparability plays a crucial role in statistical inference. In respect to the first presupposition, there is indeed a relationship between randomization and comparability. I discuss this relationship in Section 3.2. Regarding the second presupposition, the view I defend is that, although comparability is desirable for practical reasons, it is not a requirement for valid statistical inference. I present my arguments in Section 3.3.

3.2 The link between randomization and comparability

While RCT advocates often give the impression that randomization offers a quasi-certain guarantee of comparability, when pressed on the issue, they ultimately concede that randomization is ineffective for small groups and that even in the case of large groups the epistemic 'guarantee' of comparability is more along the lines of 'likely, but not infallible.' Critics take issue with these qualifications yet, they too, if pressed, cannot but acknowledge that one round of randomization can enhance comparability. The reason for these partial concessions is that randomization ensures that test and control groups are not incomparable in virtue of a biased method preferentially allocating patients having certain characteristics to one group rather than the other, but cannot guarantee that the two groups will not happen to be incomparable by chance. Since random error decreases as sample size increases, it further follows that larger groups are more likely to be comparable. This is illustrated in panels F and G of Figure 1. The relative scatter of the rerandomization distribution plots indicates that outcome variability decreases as group size increases from 100 to 1000. Lesser variability translates into a higher probability of generating comparable groups, thus providing a legitimate rationale for relying on randomization as a method for enhancing outcome comparability even when an experiment involves a single round of random allocation. At the same time, it would be unwise to ignore the fact that, for small groups, the probability that randomization generates incomparable groups is extremely high (panel E).

This much acknowledged, critics of randomization further point out that randomization does nothing for ensuring causal comparability. Howson and Urbach (2006, 196) are keen to remark that "the probability of a substantial imbalance on some prognostic factor might, for all we know, be quite large," implying that this is a serious problem for statistical inference. In the same

vein, Worrall repeatedly suggests that valid statistical inference requires comparable groups balanced in respect to each confounder, and not only for their overall effect on the outcome, something which is better achieved by experimental practices aiming to homogenize groups, such as matching and stratification:

> "Once it is accepted that for any real randomized allocation known factors might be unbalanced […] then it seems difficult to deny that a properly matched experimental and control group is better, so far as preventing known confounders from producing a misleading outcome, than leaving it to the happenstance of the tosses" (2007b, 481).

The tacit assumption here is that valid statistical inference requires the same kind of comparability needed for non-statistical causal inference, namely a one-by-one matching of each possible confounder. However, as illustrated in Figure 1, randomization can only reduce the variability–or, if we prefer, enhance the comparability–of the outcome. It says nothing about comparability in respect to causes. Assuming a deterministic relationship between causes and outcomes, differences in outcome indicate that groups differ in terms of causal factors relevant to those outcomes. However, since more than one combination of causes can lead to the same outcome, the fact that large samples are likely to be comparable in terms of outcomes does not entail that they are also comparable in terms of causal factors determining those outcomes. Quantitative analysis reveals that the probability that two groups are balanced in respect to every confounder depends on the number of confounders, their probability distributions and the nature of their interactions (Lindley 1982; Saint-Mont 2015). Relying on randomization to achieve causal comparability consistently requires large groups, in some cases larger than those required for achieving adequate statistical power and significance. Yet, only the latter concerns figure in the planning stages of an RCT, which never include a quantitative assessment of causal comparability.

3.3 Comparability is not required for valid statistical inference

The above considerations raise reasonable doubts about the role of randomization as a method for achieving the kind of comparability required by non-statistical causal inference. But is this really a problem for statistical inference? I think both advocates and critics of randomization are mistaken in assuming that comparability, either in terms of outcomes or causes, is required for valid inference of causation by statistical methods.

For one thing, physically implementing a random allocation procedure is not required for conducting a statistical test (Feinstein 1983). In the example discussed in Section 2.2, a list of outcomes for each patient is available (Figure 1, panel A), meaning that it is possible to repeatedly divide this list at random into two groups and calculate the probability of obtaining differences in outcome by chance alone. There is no need to physically allocate subjects to test and control groups.

A second argument is that, in frequentist statistics, inference of any kind hinges on the rejection of a chance explanation (the null hypothesis) according to which data scatters in the way it does because of physical processes such as those underlying measurement uncertainty, or because experimental procedures such as sampling or allocation generate an unequal distribution of confounders from one data collection context to another. A statistical model then specifies how these processes determine the scattering of the data. Finally, a statistical test assesses the probability of obtaining certain variations in data given a particular statistical model (Burnham and Anderson 2002).

In the case of a traditional RCT, the scattering of the data refers to the observed differences in outcome between test and control groups, while the null hypothesis attributes these differences to the way in which the groups were generated, namely the allocation procedure whereby a convenience sample is divided into a test and a control group. As exemplified in the case of Fisher's test of independence, the null hypothesis is ruled out when the probability of obtaining variations in data of the same magnitude as those observed in the experiment is relatively low. Conversely, if the chance explanation cannot be ruled out, the experiment cannot conclusively demonstrate the causal relevance of the treatment since the mere fact of dividing a sample in two might have sufficed to generate the observed differences in outcome.

The fact that a chance explanation involving an unequal distribution of confounders is considered and needs to be ruled out shows beyond any doubt that the causal inference deployed in the context of an RCT works explicitly on the expectation that there is a non-zero probability of generating incomparable groups. This conclusion is reinforced by the fact that the satisfaction of the comparability requirement entails that a chance explanation is impossible. If patients are comparable in respect to all confounders relevant to an outcome, then exchanging a patient from the test group with one from the control group will not make any difference to the outcome. Under these circumstances, statistical inference collapses into a version of Mill's method of difference:

given a zero probability of generating incomparable groups, if differences in outcome are observed, these differences must have been caused by something other than the process by which the groups were generated, and that irrespective of the magnitude of the differences in outcome and the size of the groups. This shows that there is no method of inference requiring both comparability and a statistical test to rule out a chance explanation, but two distinct methods requiring either one or the other.

I can foresee two objections to the above conclusion. One may be to point out that comparability concerns are omnipresent in clinical research, including RCT experimental design. Some degree of homogenization is achieved by imposing patient eligibility criteria aiming, among other things, to control for potent confounders such as comorbidities and age. Other homogenization practises include stratification, which can be used to demonstrate treatment efficacy for subclasses of patients and, if certain assumptions about the mechanism and progression of disease are met (e.g., chronic conditions), crossover designs in which the same patients are involved in both test and control conditions. It may therefore be argued that even though none of these strategies are known or expected to ensure the same level of homogenization achieved in basic science, their goal is nevertheless to ensure the highest degree of comparability possible under unfavourable circumstances.

I think this justification is incorrect. Eligibility criteria are not required for valid statistical inference. As discussed in Section 2.2, a researcher has the choice to conduct an explanatory or a pragmatic RCT, yet the statistical testing and the causal inference remain the same. The only thing that changes is the reference population: an explanatory trial assesses efficacy in respect to a specific dysfunction for patients aged 18-60, while a pragmatic trial assesses efficacy irrespective of age and comorbidities. If there is a concern for homogenization in clinical research, this is not in order to approximate the more stringent causal comparability required by non-statistical causal inference, but rather to reduce outcome variability among trial participants. From a practical point of view, this means that smaller differences in outcome can be shown to be significant or, alternatively, statistical significance can be achieved with smaller groups.

A second reply may be that statistical inference is routinely employed in conjunction with a probabilistic version of Mill's method of difference. Let us assume that comparable test and control groups can be systematically generated, yet the differences in outcome between the two

groups vary from one iteration of the experiment to the next (i.e., results are not exactly reproduced). The fact that there are differences in outcome between comparable groups suggests causation. However, a statistical test is ultimately needed in order to establish whether these differences reflect a genuine causal difference between the groups and not some chance variation linked to some source of experimental error.

My response is that the objection plays on an ambiguity about what counts as 'chance variation.' In the case of an RCT, the null hypothesis targeted by the statistical test refers to the possibility that, irrespective of whether a treatment is administered, differences in outcome are generated in virtue of the fact that a heterogenous sample of patients is divided in two groups. If the groups are comparable, then we already know that the observed variations have nothing to do with imbalances between the groups. If a statistical test is intended to assess this kind of experimental error, then this is a misuse of the test. We must look elsewhere for an explanation of the variation in outcome. Perhaps the intervention is not accurate, or variations occur in virtue of measurement uncertainty, or we are dealing with something more exotic, such as causal factors that determine the outcome in virtue of physical processes governed by probabilistic laws. Wherever the truth may lie, statistical inference will not get us any closer to it unless we first model data variation according to some hypothesis about the physical processes in virtue of which data is generated.

3.4 Why randomize?

If randomization neither ensures comparability, as required by non-statistical causal inference, nor is necessary for valid statistical inference, then why randomize? The answer lies in the fact that a statistical test only assesses the probability of a chance explanation. If the latter is deemed improbable, we are entitled to conclude that, in addition to random variation, something else is also contributing to the observed differences in outcome. By itself, this tells us nothing about the identity of this additional cause. RCTs, on the other hand, are meant to test the causal efficacy of a treatment. This raises the legitimate concern that the mere falsification of the null hypothesis does not demonstrate the causal efficacy of the treatment (Worrall 2007a, 998).

The gap between the statistical verdict and the desired conclusion is bridged by implementing statistical inference in the context of a controlled experiment. If an additional causal contribution is suggested by statistical testing, this cause can be identified as the treatment administered to

the test group in as much as the experiment did not introduce any additional causes of variation beyond those already present in the convenience sample, with the sole exception of the treatment intervention in the test group. This does not require comparable test and control groups. Instead, what is required is an experimental design ensuring that test patients are not inadvertently exposed to additional interventions or to an intervention inadvertently targeting multiple variables.

Just like in basic research, standardization and operationalization of the treatment and outcome assessment procedures, along with a close monitoring of the subjects play an important role in ensuring the accuracy of the intervention. Various forms of blinding further limit the possibility that patients and researchers willingly or unwillingly influence the outcomes of the trial. Finally, since the experimental intervention in an RCT involves the allocation of patients to test and control groups, there is an additional and very important worry to be addressed: the allocation procedure may be biased due to a common cause mechanism whereby the allocator preferentially associates treatment or control conditions with other features of the patients, such as age, lifestyle or socio-economic status. The clinical literature repeatedly emphasizes that allocation bias should be dealt with by randomization, understood as the experimental practice of random allocation. Matching and other methods associated with quasi-experimental designs, as well as passive allocation designs (natural experiments, leaving allocation to patients or their physicians), should be avoided as they may be biased by unsuspected confounders and inefficient blinding (Chalmers et al. 1983). This is why clinicians reject Worrall's recommendation [(2007b, 481); quoted in Section 3.1] that matching should be preferred to rerandomization.

Unfortunately, there is some confusion about how randomization is meant to address the threat of biased allocation. There can be an opportunity for allocation bias only if the biological systems allocated to test and control conditions don't satisfy the comparability requirement in the first place. Thus, two strategies may be adopted, one geared towards generating comparable biological systems, the other towards alleviating the effects of bias given a lack of comparability. The first strategy may be implemented by generating homogeneous populations of biological systems, the second by breaking the common cause mechanisms responsible for biased allocation. The two strategies are conflated if randomization is viewed as an effective method for implementing both strategies. If randomization generates comparable groups, this blocks the potential for allocation bias in the first place. However, the weakness of this approach is that an allowance must be made

for the possibility that comparability is not achieved. According to a second, more robust argument, randomization disrupts biasing mechanisms by randomly setting the value of the variable targeted by the intervention, effectively breaking any systematic associations with other variables (Altman and Bland 1999; Papineau 1994; Pearl 2000, 262-63, 348; Woodward 2003, 98-99, 339-42). Unlike comparability, this benefit of randomization does not depend on sample size. In this second role, randomization only ensures that there are no systematic associations between treatment and confounders. Confounders can still be associated with the test condition by chance and this may result in differences in outcome. In other words, randomization only ensures that the allocation intervention does not introduce additional confounding, in this case as systematic error due to biased allocation, over and above the unavoidable random error expected in virtue of the inherent heterogeneity of the objects of study.

According to the above argument, blinding and unbiased allocation are two different virtues, although in practice random allocation is commonly used as a device for ensuring both. Unbiased intervention is achieved by randomly setting the value of the independent variable, thus 'breaking' any potential associations with other variables. This refers to the state of the system (the experimental setup) under investigation, not the epistemic status of patients, researchers and their ability to influence the outcome of the experiment. Blinding patients and researchers at various stages of the trial, from allocation to data analysis, further removes other sources of bias.

If the accuracy of the intervention is not demonstrated, it cannot be concluded that differences in outcome are due to the treatment and its intended biological targets. This does not affect the validity of the statistical test, which only tells us how likely it is that differences in outcome could have occurred in virtue of sampling or allocation alone. Nor does it invalidate the inference that the allocation intervention is causally efficacious. It is still possible to demonstrate that doing something promotes a certain outcome even if it is not clear what exactly the intervention targets and why it works. From an epistemic point of view, the problem is not that causality cannot be inferred, but rather that causal relevance may be erroneously attributed to variables which are causally irrelevant, leading to spurious explanations and counterproductive medical practices. For example, a clinical study may show that the allocation of patients to a test group is efficacious, but erroneously attribute efficacy to the treatment and the changes in the biological targets of the treatment, when in fact the differences in outcome are due to the fact that younger patients were preferentially allocated to the test group or to the mere fact that study participants interact with health

professionals (a placebo effect). For as long as this caveat is properly understood, a non-randomized study can still provide useful information. For instance, failure to achieve a statistically significant result is a good indication that there is no genuine correlation between two variables, hence further research is not worth pursuing. A statistically significant result, on the other hand, indicates that something potentially interesting is going on, but better controlled experiments are needed in order to rule out systematic error and correctly identify the causal variables responsible for the differences in outcome (Winch and Campbell 1969).

4. Conclusion

The thesis defended in this paper is that there are two distinct methods for inferring causation in biomedical research, each requiring its own, equally distinct experimental design. These are the common, or non-statistical, inference associated with controlled experiments in basic science; and the statistical inference associated with RCTs in clinical research. I argued that when comparability of the test and control arms of an experiment can be assumed, causation can be validly inferred without any further reliance on statistical testing. Conversely, the purpose of statistical testing is to provide a method for inferring causation when one expects that incomparable groups are generated. Thus, the main difference between the two methods hinges on the satisfaction of the comparability requirement, which is in turn dictated by the nature of the objects of study, namely homogeneous vs. heterogeneous populations of biological systems. Researchers opt for a method or the other, implementing the corresponding experimental design, depending on whether there is evidence indicating that a population is homogeneous or not.

This conclusion goes against views implying that the comparability requirement must be satisfied irrespective of whether researchers work with homogeneous or heterogeneous populations, employ the common or the statistical version of causal inference, and conduct controlled experiments or RCTs. Appealing to the methodological requirements associated with one method of causal inference to evaluate experimental designs meant to support a different method of inference can only result in erroneous methodological recommendations and an improper assessment of experimental results. In particular, the claim that RCTs fail to provide better evidence for causation than non-randomized studies stems from a misunderstanding of the role of randomization in experimental design. Critics assume that causal inference in clinical research requires compara-

ble groups and that the main purpose of random allocation is to achieve comparability. Since randomization cannot guarantee comparability, the natural conclusion is that randomized studies are not superior to non-randomized ones. Yet the argument rests on false assumptions. The extent to which randomization ensures or fails to ensure comparability has no bearing on the validity of statistical inference. After all, the explicit goal of clinical research is precisely to assess causal relevance in heterogeneous populations–that is, populations known to consist of incomparable individuals. Randomization is a key requirement not because its limited capacity to enhance comparability, but because it ensures that the allocation intervention does not introduce systematic error over and above the unavoidable random error expected in virtue of the inherent heterogeneity of the population under investigation. Only if this requirement is satisfied can it be concluded that differences in outcome between test and control groups unlikely to arise by chance alone are due to the treatment administered to the test group and not to some other source of systematic error.

Bibliography

Altman, D. G., and J. M. Bland. 1999. "Treatment Allocation in Controlled Trials: Why Randomise." *British Medical Journal* 318:1209.

Ankeny, R., and S. Leonelli. 2011. "What's so Special about Model Organisms?" *Studies in History and Philosophy of Science* 42:313–23.

Baetu, T. M. 2016. "The 'Big Picture': The Problem of Extrapolation in Basic Research." *British Journal for the Philosophy of Science* 67 (4):941-64.

Burnham, K. P., and D.R. Anderson. 2002. *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*. 2nd ed. New York: Springer.

Cartwright, N. 2010. "What Are Randomised Controlled Trials Food for?" *Philosophical Studies* 147:59-70.

Chalmers, T. C., P. Celano, H. S. Sacks, and H. Smith. 1983. "Bias in Treatment Assignment in Controlled Clinical Trials." *New England Journal of Medicine* 309 (22):1359-61.

Chow, S. L. 2010. "Experimental Design." In *Encyclopedia of Research Design*, ed. N. J. Salkind. Thousand Oaks, CA: SAGE.

Clarke, A. , and J. Fujimura. 1992. *The Right Tools for the Job: At Work in Twentieth Century Life Sciences*. Princeton: Princeton University Press.

Feinstein, A. R. 1983. "An Additional Basic Science for Clinical Medicine III. The Challenges of Comparison and Measurement." *Annals of Internal Medicine* 99:705-12.

Fisher, R. A. 1947. *The Design ol Experiments*. Fourth edition ed. Edinburgh: Oliver and Boyd.

Godwin, M., L. Ruhland, I. Casson, S. MacDonald, D. Delva, R. Birtwhistle, M. Lam, and R. Seguin. 2003. "Pragmatic Controlled Clinical Trials in Irimary Care: The Struggle between External and Internal Validity." *BMC Medical Research Methodology* 3 (28):doi:10.1186/471-2288-3-28.

Hill, A. B. 1952. "The Clinical Trial." *New England Journal of Medicine* 247:113-19.

———. 1955. *Principles of Medical Statistics*. 6th ed. New York: Oxford University Press.

Howson, C., and P. Urbach. 2006. *Scientific Reasoning, a Bayesian Approach*. 3rd ed. Chicago: Open Court.
Kabisch, M., C. Ruckes, M. Seibert-Grafe, and M. Blettner. 2011. "Randomized Controlled Trials: Part 17 of a Series on Evaluation of Scientific Publications." *Deutsches Ärzteblatt International* 108 (39):663-68.
La Caze, A. 2013. "Why Randomised Interventional Studies." *Journal of Medicine and Philosophy* 38 (4):352-68.
Leighton, J. P. 2010. "Internal Validity." In *Encyclopedia of Research Design*, ed. N. J. Salkind. Thousand Oaks, CA: SAGE.
Lindley, D. V. 1982. "The Role of Randomization in Inference." *Philosophy of Science Association* 2:431-46.
Mill, J. S. . 1843. *A System of Logic, Ratiocinative and Inductive*. London: John W. Parker.
Müller-Wille, S. 2007. "Hybrids, Pure Cultures, and Pure Lines: From Nineteenth-Century Biology to Twentieth-Century Genetics." *Studies in History and Philosophy of Biological and Biomedical Sciences* 38 (4):796–806.
Papineau, D. 1994. "The Virtues of Randomization." *British Journal for the Philosophy of Science* 45:437-50.
Pearl, J. 2000. *Causality. Models, Reasoning, and Inference*. Cambridge: Cambridge University Press.
Saint-Mont, U. 2015. "Randomization Does Not Help Much, Comparability Does." *PLoS ONE* 10 (7):e0132102.
Spirtes, P., C. Glymour, and R. Scheines. 1993. *Causation, Prediction and Search*. New York: Springer-Verlag.
Steel, D. 2007. *Across the Boundaries: Extrapolation in Biology and Social Science*. Oxford: Oxford University Press.
The Cochrane Collaboration. 2011. "Cochrane Handbook for Systematic Reviews of Interventions Version 5.1.0 [updated March 2011]", in J. P. T. Higgins and S. Green (eds.), Available from www.handbook.cochrane.org.
Urbach, P. 1985. "Randomization and the Design of Experiments." *Philosophy of Science* 52:256-73.
———. 1993. "The Value of Randomization and Control in Clinical Trials." *Statistics in Medicine* 12:1421-31.
Winch, R. F., and D. T. Campbell. 1969. "Proof? No. Evidence? Yes. The Significance of Tests of Significance." *The American Sociologist* 4 (2):140-43.
Witteveen, J. forthcoming. "Regression Explanation and Statistical Autonomy." *Biology & Philosophy*.
Woodward, J. 2003. *Making Things Happen: A Theory of Causal Explanation*. Oxford: Oxford University Press.
Worrall, J. 2007a. "Evidence in Medicine and Evidence-Based Medicine." *Philosophy Compass* 2 (6):981-1022.
———. 2007b. "Why There's No Cause to Randomize." *British Journal for the Philosophy of Science* 58:451-88.