# Chapter 9
# A Millian Look at the Logic of Clinical Trials


**Roberto Festa, Gustavo Cevolani, and Luca Tambolo**

**Abstract** The use of a certain drug or treatment in the cure of a disease or condition should be based on appropriate tests of the hypothesis that said drug or treatment is efficacious in the cure of the disease or condition at hand. In this paper, we aim at elucidating how such "efficacy hypotheses" are tested and evaluated, especially in clinical trials. More precisely, we shall suggest that the principles governing the assessment of efficacy hypotheses, and, more generally, hypotheses of statistical causality, are provided by an appropriate statistical version of such a venerable procedure as the method of difference put forward by John Stuart Mill.

**Keywords** Mill's methods · Method of difference · RCT · Clinical trials · Placebo effect · Causal statistical hypothesis · Efficacy hypothesis · Cognitive structures · Scientific commonsense

## 9.1 Introduction

The use of a certain drug or treatment in the cure of a disease or condition should be based on appropriate tests of the hypothesis that said drug or treatment is efficacious in the cure of the disease or condition at hand. Let us call hypotheses of this kind "efficacy hypotheses". The problem of how to assess efficacy hypotheses is crucial not only in pharmacology and clinical epidemiology, but also in many other

R. Festa · L. Tambolo
Department of Humanistic Studies, University of Trieste, Trieste, Italy

G. Cevolani (✉)
IMT School for Advanced Studies Lucca, Lucca, Italy

Center for Logic, Language, and Cognition, University of Turin, Turin, Italy



A. LaCaze, B. Osimani (eds.), *Uncertainty in Pharmacology*, Boston Studies in the Philosophy and History of Science 338,
https://doi.org/10.1007/978-3-030-29179-2_9

fields, ranging from agriculture—where, to mention but one example, one needs to appraise the efficacy of new pesticides—to the implementation of any policy recommendation, whose efficacy needs to be properly evaluated. The principles governing the assessment of efficacy hypotheses are arguably the same in all these different areas. It then comes as no surprise that the problem of how to precisely characterize such principles has steadily attracted the attention of both practicing scientists and philosophers interested in the analysis of scientific reasoning and inference.

The great amount of work devoted to the above problem has so far produced some agreement on few crucial points. For instance, most scholars would agree that efficacy hypotheses are a special kind of causal statistical hypotheses, i.e., hypotheses aimed at identifying the cause of a statistical difference between two relevant groups (consider, as an example, the administering of a certain drug as the cause of a greater frequency of remissions among the members of the treatment group as compared to those of the control group in a properly conducted clinical trial). Granted this point of agreement, however, heated controversies concern, among other issues, the very meaning of efficacy hypotheses: philosophers and epidemiologists disagree, for instance, on the precise nature of the causal claims expressed by causal statistical hypotheses (see, e.g., Broadbent 2013). Still other debates concern the most adequate statistical methods to test such hypotheses, and many different proposals have been put forward within the two most important approaches to statistical inferences, the orthodox and the Bayesian one.

In this paper, we shall not delve into such debates. In particular, we shall not try to compare and evaluate the different methods of hypotheses assessment proposed in the literature. Instead, our aim will be to identify the common, qualitative "cognitive structures" (Kuipers 2001) underlying all of those methods, and hence to capture the "core" logic of efficacy hypotheses assessment. More specifically, we shall suggest that the basic principles governing the assessment of efficacy hypotheses, and of causal statistical hypotheses in general, are provided by an appropriate statistical version of such a time-honored procedure as the method of difference put forward by John Stuart Mill. To use Worrall's (2007a) apt phrase, our analysis will then provide a description of a key component of the "scientific commonsense", that is, a set of qualitative principles underlying such diverse theories of scientific method as Popperian falsificationism and Bayesianism, as well as the best scientific practice.

Roughly, Mill's method of difference (MD) aims at identifying the sufficient cause of a given phenomenon by eliminating all the other possible causes. The statistical version of Mill's method of difference (SMD), that we shall articulate here, aims at identifying the cause responsible for a significant statistical difference between two samples drawn from a given population by ruling out the alternative hypothesis that such difference is the result of chance. As we shall argue, SMD underlies various kinds of clinical trials and, more generally, methods of controlled inquiry widely deployed in the empirical sciences. Interestingly (as we shall see in Sect. 9.4), both philosophers and epidemiologists have often highlighted before, if only occasionally, the possible relevance of Mill's method for the assessment of causal statistical hypotheses. Still, quite surprisingly, to the best of our knowledge

none of them has provided a detailed characterization of such method as applied in the statistical case.

We shall proceed as follows. In Sect. 9.2, we shall briefly survey the key features of randomized controlled trials (RCTs), which in spite of increasing criticism are still the method of choice for the assessment of efficacy hypotheses in clinical research. We shall focus on RCTs as the clearest instance of controlled inquiry, illustrating some crucial issues concerning efficacy assessment that are common to many other methods, like case-control and cohort studies. Our conclusions, in any case, apply more generally to all kinds of controlled trials, both inside and outside epidemiology. In Sect. 9.3, we shall put forward our Millian analysis of the assessment of efficacy hypotheses. Section 9.3.1 presents Mill's method of difference; we then introduce its statistical version in Sect. 9.3.2 and show how it applies to reconstructing the logic of clinical trials in Sect. 9.3.3. As we shall argue in Sect. 9.4, the Millian insight underlying our statistical method of difference uncovers a fundamental principle of scientific commonsense. In Sect. 9.5, we shall offer some concluding remarks and hints at future research.

## 9.2 Clinical Trials

In clinical research, one typically evaluates hypotheses concerning the efficacy of treatments on the basis of *controlled trials*. We can summarize the standard aspects of controlled clinical trials as follows. In order to assess the efficacy of some treatment $T$ for a given disease or condition, researchers will go through the following fairly standard steps (see, e.g., Chow and Liu 2004; Day 2007):

(1) An experimental population $P$ of patients suffering from the disease, and eligible for the trial, is identified. ("Eligible" means here, for instance, that members of $P$ do not suffer from other pathologies, are not too old or too young, and so on.)
(2) A large sample $S$ of patients, who agreed to participate in the trial, is drawn from population $P$.
(3) $S$ is divided into two groups: the "treatment group", including the participants administered with $T$, and the "control group", including those receiving some control treatment, which may be a placebo, the standard treatment for the relevant disease, or even nothing at all.
(4) The outcome of the trial, i.e., the course of the disease in all participants, is ascertained by standardized procedures.
(5) The efficacy of $T$ is assessed on the basis of an appropriate statistical analysis of the observed outcome of the trial.

Well-known problems arise with the procedure outlined above, especially in relation to steps 3–5. A crucial issue concerns step 3, i.e., how to split the sample in the treatment and control groups. In principle, one would require the two groups to be essentially similar with respect to *every relevant variable* which might affect the course of the disease, as, for instance, sex, age, and so on. The reason is clear: the possible differences in the outcomes within the two groups can be reasonably

attributed to the action of *T* only if all the relevant variables are similarly distributed across the treatment and control groups.

However, a short reflection suggests that the above requirement is too strong in practice. Indeed, researchers can never guarantee that the treatment and control groups are essentially similar with respect to *every* relevant variable. At best, they will be able to control for those variables already known to be relevant for the course of the disease. This means that one can never exclude that the two groups will actually differ with respect to some variables, whose relevance is however unknown. In order to minimize the probability that this occurs, researchers commonly employ so called *randomized controlled trials* (RCTs),[1] that is, a kind of controlled trial where members of the relevant sample *S* are allocated to either the treatment or the control group by randomization, i.e., by using some random mechanism.

To put it differently, one crucial reason to employ RCTs is to reduce the so called selection bias, i.e., to minimize the probability that the treatment and control groups are unbalanced with respect to some relevant variable: randomization should maximize the probability that the two groups are essentially similar with respect to *any* relevant variable, including those which one does not know are relevant for the given disease. For instance, suppose that some food habit, mistakenly believed to be irrelevant for the evolution of the condition, is in fact relevant. Randomization, its advocates argue, makes it highly probable that the treatment and control groups will include approximately equal proportions of participants with the relevant habit.

The selection bias is not the only systematic error which may affect the outcome of clinical trials. Another source of biases is related to the participants' expectations concerning the effect of the treatment. In particular, knowing to which group one belongs may significantly affect the course of a disease. Moreover, also the expectations of the researchers may well distort the outcome of the trial. For these reasons, one should make sure that neither participants nor researchers will be able to learn to which group each participant has been allocated. To this aim, the standard technique is to design a *double-blind RCT*, i.e., an RCT where both participants and researchers are unaware of whether any given participant belongs to the treatment or the control group. This implies, in particular, that the patients should be unable to tell the studied treatment *T* from the treatment administered to the control group (see Jenkins 2004, Ch. 2, for a vivid illustration of the difficulties that this engenders).

In this connection, it seems beyond doubt that in the case of various diseases a treatment *T*, for the simple fact of being administered, will improve the condition of the patient; in other words, that *T* will have a more or less significant *placebo*

[1] Since the second half of twentieth century, RCTs have progressively gained the status of the method of choice first in clinical research and other fields of health care and later also in social and public policy. Their increasing popularity has recently attracted renewed critical attention on RCTs, and a growing body of literature has highlighted their limitations (see, e.g., Rothwell 2005, 2008; Bluhm 2007; Cartwright 2007, 2012; Howick 2011). Moreover, the alleged advantages deriving from randomization have been doubted (see, e.g., Urbach 1985; Howson and Urbach 2006. Ch. 6; Worrall 2007b; Howick 2011; Saint-Mont 2015), and a vocal minority has advocated instead Bayesian trials (see, e.g., Kadane 1996; Teira 2011).

*effect* (see, e.g., Benedetti 2008). For the moment, a placebo may be defined as any treatment that the practitioner believes to have no specific therapeutic efficacy for a certain disease.[2] A classic example of placebo is the *pillula panis* (bread pill), widely used in medicine since at least the Eighteenth century. Given that as hinted above the placebo effect manifests itself in connection with various diseases, assessing the efficacy of a treatment $T$ for such conditions typically means to check whether the effect of $T$ is merely a placebo effect. This is the reason why the members of the control group are usually administered an appropriate placebo, which is as similar as possible to $T$. This is crucial since the intensity of the placebo effect is affected by many variables, including the mode of administration and the context of medical intervention. As an example, generally a treatment administered at the patient's home does not have the same effect as when administered at the hospital (see Kaptchuk et al. 2000). For this reason, researchers testing a new treatment $T$ look for a placebo such that its outlook, its mode of administration, and its collateral effects are identical to those of $T$. If this condition is satisfied, then the possible differences between the course of the disease in the treatment and control groups may be attributed with some confidence to the active component of the treatment. What this exactly means will be the topic of the next section.

## 9.3 Assessing Hypotheses with the Method of Difference

As recalled in the foregoing section, a clinical trial aims at assessing the efficacy of some treatment for a given condition. In this section, we shall argue that the logical structure of clinical trials is best reconstructed as a statistical version of the well-known "method of difference", an inference schema introduced by John Stuart Mill (1806–1873) in his *System of Logic* (1843). We shall start by briefly describing how the method of difference may be used in assessing hypotheses of sufficient causality (Sect. 9.3.1). We shall then introduce a statistical version of this method that, as we shall argue, may be applied to the assessment of hypotheses of statistical causality (Sect. 9.3.2) and, in particular, of hypotheses of efficacy as tested in clinical trials (Sect. 9.3.3).

[2] A more precise definition will follow in Sect. 9.3.3. There is a wide scientific and philosophical literature devoted to the attempt to provide an adequate definition of the concept of placebo: see, for instance, Grünbaum 1993, Kaptchuk 1998, Macedo et al. 2003, Howick 2011, Broadbent 2013, Miller et al. 2013, Holman 2015.

### 9.3.1 Hypotheses of Sufficient Causality

Imagine that a group of researchers is investigating the causal relations occurring among some features exhibited by the individuals[3] belonging to a given population $P$. In the course of their inquiry, researchers might come to consider what we call hypotheses of sufficient causality. A hypothesis of sufficient causality $H$ has the form "$C$ is a sufficient cause of $E$ (in $P$)", where both $C$ and $E$ are features that may characterize the individuals of $P$. It is common to construe $H$ as a universal hypothesis, restating it as follows:

> All the individuals (of $P$) exhibiting $C$ also exhibit $E$,

or, equivalently, "If an individual (of $P$) exhibits $C$, it also exhibits $E$". In the following, we shall often omit the reference to the relevant population $P$, which we take as tacitly understood. Moreover, we shall call "phenomenon" any feature $E$ which is investigated as the possible effect of some cause, and "conditions" the features $C$, $C_1, \ldots, C_n$ which are considered as possible sufficient causes of $E$. To illustrate, suppose that researchers are investigating a case of foodborne illness, so that the relevant phenomenon $E$ is some kind of intoxication. The aim of the investigation is then to discover a sufficient cause of $E$ among a number of possible conditions $C$, $C_1, \ldots, C_n$, like having consumed particular kinds of food (an example will follow in moment).

Dealing with the inference of causal hypotheses, Mill proposed his famous five "canons" of induction, which included the *method of difference* (Mill 1843, book III, ch. VIII, sec. 2, p. 256):

> if an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring in the former; the circumstance in which alone the two instances differ, is the effect, or the cause, or an indispensable part of the cause, of the phenomenon.

Since Mill's formulation may not be fully transparent, in the literature one finds various ways of restating the method of difference in more rigorous terms (see van Heuveln 2000 for a survey). Our proposed reconstruction of the method is the following:

(MD) Suppose you are investigating a phenomenon $E$ (in population $P$) and accept the assumption $A$ that "At least one of the conditions $C$, $C_1, \ldots, C_n$ is a sufficient cause of $E$". Moreover, suppose that, by considering two different individuals $a$ and $b$ in $P$, you make the observation $O$ that "$E$ is present in $a$ and absent in $b$, $C$ is present in $a$ and absent in $b$, and $C_1, \ldots, C_n$ are all present both in $a$ and in $b$". Then you can infer $H$, i.e., that $C$ is a sufficient cause of $E$.

[3] We use here the term "individual" in a quite wide sense here, to denote not only persons or organisms, but also different kinds of objects or events. For instance, different portions of lands, or even different policies, may be individuals in our sense (cf., e.g., Holland 1986, p. 945).

Let us illustrate how MD works by means of a simple example. Suppose that Adam and Eve are two friends who had dinner together last night. In the morning, Adam feels sick, while Eve is fine: thus, there is a phenomenon $E$ (the illness in question, or better its symptoms, like, say, vomit, aches, and fever) affecting Adam but not Eve. A reasonable assumption ($A$) is that Adam's illness may depend on some of the foods served at dinner being noxious for Adam. After a short reflection, the two friends realize that only Adam, but not Eve, had mushrooms as garnish for the steak, while all other courses — say soup, steak, and cake — were the same for both. In other words, there is a condition $C$ (having eaten mushrooms) which affects Adam but not Eve, whereas all other conditions — i.e., $C_1$ (eating soup), $C_2$ (eating steak), and $C_3$ (eating cake) — have been common for both of them. Such observation ($O$) is graphically represented in Table 9.1. On the basis of it, one has reasons to suspect that eating mushrooms ($C$) was the cause of Adam's intoxication. Moreover, one can presume that also other guests of the same restaurant, who also had mushrooms for dinner last night, will be probably intoxicated as well.

Let us now consider the general case. Table 9.2 represents the general form of observation $O$ appearing in our formulation of the method of difference. Here the presence of $E$ in $a$ is indicated by "+" and its absence in $b$ by "−"; similarly, "+" ("−") is used to indicate the presence (absence) of each condition $C$, $C_1, \ldots, C_n$ in $a$ and $b$.

It should be clear where the name "method of difference" comes from. It refers to the key intuition underlying Mill's method: observation $O$ shows that, as far as features $E$, $C$, $C_1, \ldots, C_n$, are concerned, there are only two *differences* between the individuals $a$ and $b$: first, $C$ is present in $a$ while absent in $b$; second, $E$ is present in $a$ while absent in $b$. It is then natural to conclude that the former difference is the cause of the latter, i.e., that $a$ has $E$ just *because* it has $C$. Moreover, under appropriate circumstances, one may generalize this inference and conclude that, presumably, $C$ is a sufficient cause of $E$ in the whole population $P$.

**Table 9.1** A toy application of the method of difference MD: eating mushrooms is identified as the sufficient cause of Adam's intoxication

| Individuals | Possible sufficient causes | | | | Phenomenon |
|---|---|---|---|---|---|
| | Mushrooms | Soup | Steak | Cake | Illness |
| Adam | Yes | Yes | Yes | Yes | Yes |
| Eve | No | Yes | Yes | Yes | No |

**Table 9.2** General form of the method of difference MD: observation $O$ shows that the presence of condition $C$ in individual $a$, but not in $b$, is a cause of the presence of $E$ in $a$, but not in $b$.

| Individuals | Possible sufficient causes | | | | | | Phenomenon |
|---|---|---|---|---|---|---|---|
| | $C$ | $C_1$ | ... | $C_i$ | ... | $C_n$ | $E$ |
| $a$ | + | + | | + | | + | + |
| $b$ | − | + | | + | | + | − |

The plausibility of the inference licensed by MD is relatively uncontroversial. Still, long debates have addressed the question of how to rigorously characterize and justify such kind of inference. In particular, one may ask under what conditions MD can be properly applied, so that $H$ can be reliably inferred from observation $O$. For what concerns us here, it will be sufficient to recall the following points, which are generally agreed upon (cf. Mackie 1980; Skyrms 2000; Hurley 2012).

First, it should be clear that $H$ is a deductive consequence of $A$ & $O$. In fact, if assumption $A$ concerning the possible sufficient causes of $E$ is true, and if observation $O$ is correct, then $H$ must be true, i.e., $C$ must be a sufficient cause of $E$. In our example, if one is sure that the symptoms affecting $a$ may *only* depend on one of the foods served at dinner, then eating mushrooms has to be the cause of the illness, since all the other possible causes have been excluded. As already this simple example suggests, however, in typical situations one cannot be completely certain that $A$ is true. Consequently, one cannot be sure that $H$ is true, in spite of the fact that $H$ deductively follows from $A$ & $O$.

Second, $H$ is *not* a deductive consequence of observation $O$ alone. In fact, it is logically possible that $H$ is false even though $O$ is true: we cannot rule out that, sooner or later, one will stumble upon an individual of $P$ exhibiting $C$ but not $E$. This would mean that assumption $A$ is false, i.e., that none of the conditions $C$, $C_1$, ..., $C_n$ is actually a sufficient cause of $E$. For instance, suppose it turns out that, on the morning Adam feels sick, other guests of the restaurant, who also had mushrooms for dinner, are however perfectly fine. In such scenario, eating mushrooms can be excluded as a sufficient cause of $a$'s illness: some other cause, such as a previously acquired gastroenteritis, has to be identified.

Third, even if $H$ cannot be deductively inferred from $O$ alone, it can be however *inductively* inferred from $O$. Depending on the underlying epistemological framework, one can defend this claim in different ways. Within the Bayesian approach, for instance, one may argue that $H$ inductively follows from $O$ in the sense that $O$ increases the initial probability of $H$ or, in equivalent terms, that $O$ confirms $H$ (see, e.g., Crupi 2016). This can be proved as follows. Let $p(A)$ be the initial probability attributed to $A$ in the light of the available background knowledge. Since $H$ logically entails $A$, but not vice versa, inequality $p(A) > p(H)$ holds. However, $O$ logically implies that $H$ and $A$ are, in fact, logically equivalent. In turn, this implies that $p(H|O) = p(A|O)$. Since there is no reason to think that $O$ affects the initial probability of $A$, one can plausibly assume that $p(A|O) = p(A)$ holds. From the (in)equalities so obtained—i.e., $p(H|O) = p(A|O)$, $p(A|O) = p(A)$, and $p(A) > p(H)$—it follows that $p(H|O) > p(H)$. This means that $O$ increases the initial probability of $H$ or, which is the same, that $O$ confirms $H$. Finally, one may also argue that $H$ inductively follows from $O$ in another sense: i.e., that the final probability $p(H|O)$ of $H$ given $O$ is significantly high. Indeed, given the above equalities $p(H|O) = p(A|O)$ and $p(A|O) = p(A)$, $p(H|O)$ will be high just in case $p(A)$ is. In other words, if the initial probability attributed to $A$ on the ground of the available background knowledge is high, the final probability of $H$ will be comparably high.

In sum, while the method of difference MD, in typical circumstances, licenses conclusions that are not certainly true, it can provide good reasons to accept them with reasonably high confidence.

### *9.3.2 Hypotheses of Statistical Causality*

In most real-life circumstances, and especially in the medical domain, it is typically very difficult, if not plainly impossible, to identify a sufficient cause of a phenomenon $E$. For instance, in many cases, eating mushrooms of a certain species is not a sufficient cause of illness: it may be that most, but not all, individuals eating them will feel sick, perhaps depending on their personal allergies, intolerance, medical history, and on the complex combination of these and other, possibly unknown, factors. Even in such cases, however, the search for causal relations is not entirely hopeless. In fact, researchers will often rest content with investigating what one may call hypotheses of statistical causality. There are many ways of interpreting a hypothesis $H$ like "$C$ is a statistical cause of $E$ in population $P$"; one, particularly useful for our purposes, is the "comparative model" proposed by Giere (1979, sec. 9.5).

According to Giere, we may construe the hypothesis "$C$ is a statistical cause of $E$" as comparing two "hypothetical populations", to be called $P_C$ and $P_{\neg C}$. These are obtained by altering the actual population $P$ in two different ways: $P_C$ by imaging "everyone in the population [$P$] as having [$C$], but otherwise the same in all relevant respects" (Giere 1979, p. 177), and $P_{\neg C}$ by imaging no one in $P$ as having $C$, but otherwise the same in all relevant respects. Then, "$C$ is a statistical cause of $E$" simply means that the number of individuals exhibiting $E$ is greater in $P_C$ than in $P_{\neg C}$. For instance, suppose that $P$ is the group of people who had dinner in our restaurant last night; probably, some of them did eat mushrooms ($C$), and others did not ($\neg C$). Within this framework, $P_C$ is construed as the same as population $P$, with the only difference that everybody had mushrooms for dinner; and $P_{\neg C}$ is the same as $P$ except for the fact that nobody had mushrooms. The meaning of hypothesis $H$, in this case, is that there would be more sick people in the former population than in the latter.

In the following, we shall express a hypothesis $H$ of statistical causality in this form:

> $C$ is a *statistical cause* of $E$ in $P$ if and only if $\%E_C > \%E_{\neg C}$.

Here, $\%E_C$ and $\%E_{\neg C}$ are the percentages of the individuals exhibiting $E$ in $P_C$ and $P_{\neg C}$, respectively; in other words, $H$ says that the frequency of $E$ in $P_C$ is greater than in $P_{\neg C}$, or that $C$ is a "positively relevant factor" for $E$. This latter terminology is justified by noting that a statistical cause, in the sense just defined, is but a special case of a more general notion, that of a "causally relevant" factor. In words, $C$ is causally relevant for $E$ just in case the frequencies of $E$ in $P_C$ and $P_{\neg C}$ are not identical (i.e., $\%E_C \neq \%E_{\neg C}$). In the opposite case, when $\%E_C = \%E_{\neg C}$, $C$ is

causally *irrelevant* for $E$. Thus, a statistical cause is a particular kind of causally relevant factor, i.e., a *positively* relevant factor.

Assessing a hypothesis of statistical causality $H$, then, amounts to assessing whether $C$ is a positively relevant factor for $E$. The most direct way to do this would be to check whether the number of individuals having $E$ is greater in $P_C$ than it is in $P_{\neg C}$, i.e., whether $\%E_C > \%E_{\neg C}$. This, however, is impossible, for the plain reason that both $P_C$ and $P_{\neg C}$ are hypothetical, and not actual, populations. Since the only actual population is $P$, what one can do is to draw a large sample $S$ from $P$, and then divide the sample into two parts, $S_C$ and $S_{\neg C}$, such that all the individuals in $S_C$ have $C$ and none of the individuals in $S_{\neg C}$ has $C$ (more below on the appropriate procedures to adequately split $S$ in this way). Arguably, $S_C$ and $S_{\neg C}$ "play the role of samples from the two hypothetical populations" $P_C$ and $P_{\neg C}$, since these samples "are just *as if* they had been sampled from the two hypothetical populations" (Giere 1979, p. 249).[4] Finally, one can check whether the percentage $\%e_C$ of the individuals exhibiting $E$ in $S_C$ is higher than the percentage $\%e_{\neg C}$ of individuals having $E$ in $S_{\neg C}$.

The intuition outlined above underlies what we call the statistical method of difference, SMD for short:

(SMD) Suppose you are investigating a phenomenon $E$ (in population $P$) and you accept the assumption $A$ that "Conditions $C$, $C_1$, ..., $C_n$ include all causally relevant factors for $E$". Moreover, suppose that, by considering two samples $S_C$ and $S_{\neg C}$ from $P$, you make the observation $O$ that "$\%e_C$ is much greater than $\%e_{\neg C}$, while the two samples are otherwise 'balanced' with respect to $C_1$, ..., $C_n$". Then you can infer $H$, i.e., that $C$ is a statistical cause of $E$.

Here, that the two samples are balanced means that they include roughly the same percentage of individuals having $C_i$ as population $P$. More precisely, note that, by their very definition, the two hypothetical populations $P_C$ and $P_{\neg C}$ contain exactly the same percentage $\%C_i$ of individuals exhibiting $C_i$, for any condition $C_i$ different from $C$; as far as the latter condition is concerned, $P_C$ and $P_{\neg C}$ contain, respectively, 100% and 0% of individuals having $C$. This latter point holds also for the two

[4] Giere's passage is worth quoting at length:

> [...] we *create* samples that play the role of samples from the two hypothetical populations [$P_C$ and $P_{\neg C}$]. We do this by taking a sample from the real population and then dividing it in two parts [i.e., $S_C$ and $S_{\neg C}$]. All the individuals in one part [i.e., $S_C$] are experimentally manipulated so that they have the condition [$C$] that defines the hypothetical population [$P_C$]. The individuals in the other part [i.e., $S_{\neg C}$] are experimentally manipulated so that they have the condition [$\neg C$]. These two groups, then, are just *as if* they had been sampled from the two hypothetical populations [$P_C$ and $P_{\neg C}$].

Here Giere refers to an experimental inquiry, where researchers can control the condition $C$ by manipulating individuals taken from $P$ in order to create, as it were, $C$-individuals. However, Giere's insight can be easily adapted to purely observational contexts, where researchers lack full control on $C$, but still can observe whether an individual drawn from $P$ exhibits, or not, $C$. Also in this case, in fact, they can adequately form the two relevant samples $S_C$ and $S_{\neg C}$.

samples $S_C$ and $S_{\neg C}$: all individuals in the former, and none in the latter, exhibit $C$. What distinguishes the hypothetical populations from the actual samples is that the percentage of individuals having $C_i$ (for $C_i \neq C$) may differ in the two samples. This means two things: first, that the frequency of individuals having $C_i$ in $S_C$ may differ from the corresponding frequency in $S_{\neg C}$; and, second, that both those figures may differ from the actual frequency $\%C_i$ of the conditions in $P$. In short, the frequencies of individuals exhibiting $C_i$ in $S_C$, $S_{\neg C}$, and $P$ will often be, at best, only approximately equal; this is normally expressed by saying that $S_C$ and $S_{\neg C}$ are "balanced" with respect to any $C_i \neq C$. Table 9.3 displays the ideal special case where the two samples are "perfectly balanced".

The structural similarity between Table 9.3 and Table 9.2 makes it clear why SMD can be construed as a statistical version of MD. In fact, the key intuition underlying both methods is the same. In the present case, this amounts to the following: there are only two *substantial* differences between the two samples $S_C$ and $S_{\neg C}$. The first, as just recalled, is that, by definition, all members of $S_C$ and no member of $S_{\neg C}$ exhibit $C$. The second substantial difference, highlighted by observation $O$, is that the frequency of $E$ in $S_C$ is much higher than in $S_{\neg C}$. Hence, one can plausibly infer that the second difference depends on the first, in the sense that the higher frequency of $E$ in $S_C$ is caused by the higher frequency of $C$ in $S_C$. Moreover, under appropriate circumstances, one may generalize this inference to the whole population $P$, and conclude that $C$ is presumably a statistical cause of $E$ in $P$.

For the sake of clarity, let us slightly modify our restaurant example. Suppose that the relevant population $P$ includes some hundreds workers of a big factory, some of which suffered from a severe form of intoxication after consuming lunch at the factory's cafeteria. Suppose that the cafeteria's menu had soup, steak, mushrooms, and cake for lunch, and that these are all causal factors considered as possibly relevant for the intoxication. Now suppose that $S_C$ and $S_{\neg C}$ are two groups of 100 workers each, the former including only workers who had mushrooms for lunch, the latter including only people who did not have mushrooms for lunch. Table 9.4 displays a distribution of all possible causal factors among the members of such groups. In such a case, SMD would allow one to infer that eating mushrooms was probably a statistical cause of foodborne illness.

Our statistical version of Mill's method provides, we submit, an appropriate characterization of the logical structure, i.e., of the qualitative conceptual aspects, of a wide variety of statistical procedures deployed in empirical sciences. As we shall

**Table 9.3** General form of the statistical method of difference SMD

| Samples | Possible causally relevant factors | | | | | | Phenomenon |
|---|---|---|---|---|---|---|---|
| | $C$ | $C_1$ | ... | $C_i$ | ... | $C_n$ | $E$ |
| $S_C$ | 100 | $\%C_1$ | | $\%C_i$ | | $\%C_n$ | $\%e_C$ |
| $S_{\neg C}$ | 0 | $\%C_1$ | | $\%C_i$ | | $\%C_n$ | $\%e_{\neg C}$ |

Observation $O$ shows that the fact that all individuals in one sample, and none in the other, share condition $C$, is a cause of the difference in the proportion of individuals exhibiting $E$ in the two samples: it is assumed that $\%e_C$ is much greater than $\%e_{\neg C}$ and that the frequency of each $C_i$ in both samples is (at least approximately) equal to $\%C_i$. See Table 9.4 for an illustration

**Table 9.4** A toy application of the statistical method of difference SMD (see Table 9.3 for the general form): eating mushrooms is identified as a statistical cause of intoxication among the factory's workers

| Samples | Possible causally relevant factors | | | | Phenomenon |
|---|---|---|---|---|---|
| | Mushrooms | Soup | Steak | Cake | Intoxication |
| $S_C$ | 100/100 | 51/100 | 68/100 | 27/100 | 89/100 |
| $S_{\neg C}$ | 0/100 | 55/100 | 71/100 | 25/100 | 22/100 |

better detail in Sect. 9.4, several authors have noted that some statistical version of Mill's method of difference should be applicable in assessing hypotheses of statistical causality. Still, to the best of our knowledge, no explicit formulation of principles such as SMD is to be found in the philosophical and statistical literature. This is rather surprising, among other things, since it seems quite clear that SMD is widely, if tacitly, accepted by statisticians, philosophers, and working scientists. One reason for neglecting SMD (or similar formulations of the same principle) may be that scholars have focused on the differences, more than on the fundamental similarities, among the various statistical procedures for the assessment of hypotheses of statistical causality, which have been the topic of heated debate.

In this connection, we believe that all these procedures may be construed as alternative ways to specify and implement the central intuitions expressed by SMD. To be sure, one can construe each step in SMD in different ways, and this leads to apparently wildly diverse conceptions of the assessment of hypotheses of statistical causality. Of particular importance is the disagreement on three related points: the selection of samples $S_C$ and $S_{\neg C}$ from population $P$, the nature of the inference licensed by SMD, and the "strength" of this inference in relation to the size of the samples. Let us briefly comment on these points in turn.

*Selection of Samples $S_C$ and $S_{\neg C}$* As the reader will recall, observation $O$ in SMD requires that, for any causally relevant factor $C_i$ different from $C$, the frequencies of $C_i$ in the two samples $S_C$ and $S_{\neg C}$ are approximately equal to the frequency of $C_i$ in $P$. Typically, observing that this is indeed the case is only possible because $S_C$ and $S_{\neg C}$ are formed in advance exactly in order to meet such requirement. There are basically two possible ways of forming balanced samples, briefly described below.

Under the first, "restrictive" construal of $S_C$ and $S_{\neg C}$, conditions $C_1, \ldots, C_n$ are just the conditions that, on the basis of the available background knowledge, one knows to be causally relevant factors for $E$ in $P$ (or, more inclusively: the conditions that, given one's background knowledge, one believes are certainly or likely causally relevant). Under this construal, one typically knows, at least approximately, the frequencies $\%C_1, \ldots, \%C_n$ of the corresponding conditions in $P$. Given this restriction, one will draw $S_C$ and $S_{\neg C}$ from $P$ in such a way that the two samples will contain roughly the same percentages of individuals exhibiting each of the conditions $C_1, \ldots, C_n$.

A second, "unrestrictive" construal of $S_C$ and $S_{\neg C}$ is that conditions $C_1, \ldots, C_n$ are *all* the possible, known and unknown, causally relevant factors for $E$ in

$P$. This means that the set $C_1, \ldots, C_n$ is actually split in two parts: the class of causally relevant factors $C_1, \ldots, C_k$ which are known to be causally relevant, and the set of causally relevant factors $C_{k+1}, \ldots, C_n$ which are not known to be causally relevant. Under this construal, one cannot know, even approximately, the actual frequencies of all the conditions $C_1, \ldots, C_n$ in $P$. Hence, in order to guarantee that (very probably) $S_C$, $S_{\neg C}$, and $P$ contain approximately the same percentages of individuals exhibiting each condition, one needs to apply appropriate procedures to randomly draw the two samples from $P$. A huge literature discussing such randomizing procedures exists, and we shall not discuss this issue further.

*Inferring H from O in SMD* It is well-known that proponents of different approaches to statistical inference disagree quite strongly on the nature of the inference leading from observation $O$ to the hypothesis of statistical causality $H$. Still, they agree at least on one point, i.e., that such inference is not deductive, since $H$ follows neither from $O$ alone nor from $A$ & $O$ (contrary to what happened with MD). In turn, this means that the inference of $H$ from either $O$ or $A$ & $O$ is an inductive inference, which, in any case, cannot afford full certainty to $H$.

Granted this, heated debates concerning the most appropriate way to characterize such inductive inference divide the supporters of different versions of both the orthodox and the Bayesian approach (e.g., Romeijn 2017). Even a quick look at the various positions defended within the debate is sufficient to recognize two "families" of approaches. The first construes the inference licensed by SMD as leading to some form of "inductive acceptance", meaning that $H$ is fully, if only tentatively, believed as the conclusion of that inference.[5] The second view is that such inference assigns $H$ a (high) degree of credibility, where "credibility" is here a catchall term used to denote epistemic probabilities, p-values, indexes of significance, and other quantitative notions of credence.

*Sample Size and Inference Reliability* In all accounts of statistical inference, the question arises as to what is the "sufficient size" that samples $S_C$ and $S_{\neg C}$ are required to exhibit in order to guarantee that the inference from $O$ to $H$ is indeed reliable. The issue is intertwined with the specific interpretation of inductive inference adopted in each case, so opinions diverge also in this matter. At the very least, all accounts agree on one simple point: that *ceteris paribus*, the greater the sample, the more reliable the corresponding inference. A further point of agreement concerns another important feature of the sample, which affects the reliability of the corresponding inference. This is the following: the greater the extent to which $\%e_C$ exceeds $\%e_{\neg C}$, the stronger our confidence that $\%E_C$ is greater than $\%E_{\neg C}$. In other words: when $\%e_C$ is *much greater* than $\%e_{\neg C}$, this increases the probability that $\%E_C$ is *greater* than $\%E_{\neg C}$, i.e., that $C$ is a statistical cause of $E$. Again, granted this, opinions differ on the proper way of measuring how much $e_C$ exceeds $\%e_{\neg C}$,

[5] This notion of inductive acceptance, where acceptance may be construed in several, cognitive or pragmatic, senses, has been thoroughly analyzed by philosophers of science, starting already with Hempel (1960; see also Levi 1967 and Hilpinen 1968).

and hence how strong is the corresponding inductive inference (Broadbent 2013, Ch. 9).[6]

We conclude by emphasizing again that, however important these points of disagreement are, they should not obscure the fact that different views concerning sample size and selection, as well as the nature and strength of inductive inference, are but variations on a common logical structure, which is provided, as we argued, by SMD.

### *9.3.3 Hypotheses of Efficacy*

An "efficacy hypothesis" *H* is the attempt of identifying the cause responsible for a statistical difference between two relevant groups. In a typical case, *H* says that some drug or treatment *T* is efficacious in the cure of some disease or condition (cf. Sect. 9.2). Performing an RCT amounts to testing the corresponding claim that the administration of *T* is the cause of a greater frequency of remissions among the members of the treatment group as compared to those of the control group. In what follows, we shall argue that the statistical method of difference SMD is an adequate tool for assessing efficacy hypotheses, which are but a special kind of causal statistical hypotheses as discussed in Sect. 9.3.2. To this purpose, we shall focus on an imaginary RCT in which a given treatment *T* is administered to the treatment group and a placebo to the control group. Let us start by clarifying what a placebo is in relation to some treatment *T*.

In general, one can view a treatment *T* as a combination of two components: a *specific* or *distinctive* component *D*, believed to favor the remission of the relevant disease, and a *generic* component *G* that is constituted, as it were, by the kind of treatment to which *T* belongs (see Grünbaum 1993, Ch. 3, who calls *D* and *G*, respectively, the characteristic and the incidental factors of *T*). For instance, suppose that treatment *T* consists in administering some substance, say vitamin C, which is believed to favor quick remission of some condition, like the common cold. Suppose that the substance is administered orally: for 7 days the patient takes, every 8 h, a small, circular, white, bitter-tasting pill. Of course, the pill will contain a given amount of vitamin C; but, besides this, it will also contain certain quantities of various excipients (e.g., lactose, flavors, etc.) that, as far as we know, can in no way

[6]Interestingly, this issue seems an instance of the so called problem of measure sensitivity. In the theory of Bayesian confirmation, such label refers to the existence of a plurality of non-equivalent ways of measuring how much some piece of evidence confirms a given hypothesis, and to the fact that the soundness of theoretical arguments surrounding confirmation crucially depends on the specific measure adopted (see Fitelson 1999; Brössel 2013; Crupi 2016, sec. 3.4; Festa and Cevolani 2017). As the present case suggests, the problem may arise in virtually any area where one explicates informal concepts (like confirmation, inference strength, and the like) by using formally defined models or measures. On probabilistic measures of causal strength see Fitelson and Hitchcook (2011) and Sprenger (2018).

influence the course of the disease. We shall say that vitamin C, taken in the dose mentioned above, is the specific component $D$ of $T$, while the pills are its generic component $G$, i.e., the "vehicle" through which vitamin C is administered to the patient.

The distinction just introduced allows us to define the concept of a "placebo corresponding to $T$"—for short: a $T$-placebo. We shall say that a certain treatment is a $T$-placebo if it has the same generic component of $T$, so that it appears identical to $T$, but lacks the specific component of $T$. By designing an RCT, one aims at testing the hypothesis $H$ that $T$ is efficacious, i.e., that if patients were administered $T$, the frequency of positive outcomes would be higher than if they were administered a $T$-placebo. Based on our characterization of $H$ as a causal statistical hypothesis, we can proceed to describe the situation as follows. The relevant population $P$ is formed by all the patients suffering from a given condition, and the phenomenon $E$ under investigation is the possible remission or positive outcome recorded for any patient. Then our efficacy hypothesis concerning treatment $T$ has the following form:

$D$ is a causally positive factor for $E$ in $P$,

where, to recall, $D$ is the specific component of $T$. As we know, SMD implies that testing such a hypothesis means comparing the relative frequencies of $E$ in two balanced samples drawn from $P$. Indeed, this is exactly what an ideal RCT does. The first sample, $S_D$, corresponds to the experimental group, containing patients from $P$ who are all administered both the specific component $D$ and the generic component $G$ of the treatment. The second sample, $S_{\neg D}$, is instead the control group, where all patients receive $G$ (via the $T$-placebo), but none of them receives $D$.

Table 9.5, which is but a small variation on Table 9.3, shows how SMD is applied to the testing of the efficacy hypothesis in our imaginary RCT.

The causally relevant factors $C_1, \ldots, C_n$ appearing besides $D$ and $G$ in the table are the so called "prognostic factors", which can influence the probability of a positive outcome. Since the only substantial difference between the two samples concerns the frequency of patients treated with the specific component $D$, one can infer that this is the cause of the observed difference in the frequencies $\%e_D$ and $\%e_{\neg D}$ of remissions in the two cases, i.e., that treatment $T$ is efficacious.

**Table 9.5** General form of the statistical method of difference SMD as applied to clinical trials

| Samples | Possibly relevant causal factors | | | | | | | Phenomenon |
|---|---|---|---|---|---|---|---|---|
| | $D$ | $G$ | $C_1$ | ... | $C_i$ | ... | $C_n$ | $E$ |
| $S_D$ | 100 | 100 | $\%C_1$ | | $\%C_i$ | | $\%C_n$ | $\%e_D$ |
| $S_{\neg D}$ | 0 | 100 | $\%C_1$ | | $\%C_i$ | | $\%C_n$ | $\%e_{\neg D}$ |

Observation $O$ shows that the fact that all individuals in one sample, and none in the other, are given the specific component $D$ of the relevant treatment, is a cause of the difference in the proportion of individuals exhibiting $E$ in the two samples: it is assumed that $\%e_D$ is much greater than $\%e_{\neg D}$ and that the frequency of each $C_i$ in both samples is (at least approximately) equal to $\%C_i$. See Table 9.6 for an illustration

**Table 9.6** A toy application of the statistical method of difference SMD to an imaginary RCT (see Table 9.5 for the general form): the assumption of vitamin C is identified as a statistical cause of early remission from the common cold

| Samples | Possible causally relevant factors | | | | | Phenomenon |
|---|---|---|---|---|---|---|
| | Vitamin C | White pill | Sex | Sporting | Vegetables | Early remissions |
| $S_D$ | 100/100 | 100/100 | 48/100 | 29/100 | 27/100 | 89/100 |
| $S_{\neg D}$ | 0/100 | 100/100 | 51/100 | 32/100 | 25/100 | 22/100 |

Table 9.6 (which is structurally identical to Table 9.4) illustrates an imaginary RCT designed to test the efficacy of vitamin C in favoring the quick remission from the common cold. Both the treatment and the control group include 100 participants; all the members of the former are administered vitamin C in the form of a white pill, while all the members of the latter receive the corresponding placebo. The only other prognostic factors are sex (expressed as the number of men in each sample), doing sport and eating vegetables. Given the displayed figures, SMD would allow one to conclude that vitamin C is probably a positively relevant factor for the remission from common cold.

Table 9.6 should make it clear that, as we argue, SMD describes the qualitative structure of all kinds of RCTs. Indeed, SMD arguably describes the logical structure of any kind of controlled trial, randomized or not. In fact, to briefly sum up the central point of our discussion so far, the aim of all controlled trials is to assess hypotheses of statistical causality, i.e., to identify the likely cause of an observed difference between two sufficiently large samples of the relevant population. As we argued here, the logic governing such assessments is essentially the same underlying Mill's method of difference.

## 9.4 Mill's Method and the Cognitive Structure of Scientific Commonsense

In the previous section, we argued that most standard methods for the assessment of hypotheses of statistical causality, including RCTs and other kinds of controlled experiments widely used in the medical and social sciences, essentially rely on a statistical version of Mill's method of difference, i.e., SMD. In this section, we shall briefly survey some Millian insights in the philosophical and scientific literature, and show how the logic of controlled inquiry, as summarized in SMD, meshes well with recent reflections on the fundamentals of the scientific method.

*Millian Insights in Epidemiology and Statistics* To the best of our knowledge, a study of the influence of Mill's methods on the development of modern statistical methods, and of their reception in the philosophical, epidemiological and, more generally, scientific literature of the last 150 years is still to be done. Here we shall restrict our attention to the works of a number of philosophers, epistemologists of

medical sciences, and epidemiologists who noticed the relevance of Mill's methods for standard statistical procedures.

To begin with, let us point out that, as a matter of historical coincidence, Mill wrote the *System of Logic* during what Morabia (2004) calls the "pre-formal" phase of epidemiology. During this phase, lasting until the end of nineteenth century, "scientists used population thinking and group comparisons, spontaneously, without referring to some theory [ . . . ] The mathematical and philosophical bases existed but no formal theory [ . . . ] none of the concepts and methods had been *formally* defined" (Morabia 2004, p. 107). Some main figures of this era include such pioneers as British physicians William Farr (1807–1883) and John Snow (1813–1858), French physiologist Claude Bernard (1813–1878), and Hungarian physician Ignaz Philipp Semmelweis (1818–1865). Studying the possible relations and interactions between Mill, who employs a number of medical examples in his work (cf., e.g., Mill 1843, Book III, Ch. IX, §§ 1, 6, 7), and the epidemiologists of his time would be of great interest, especially in order to establish whether Mill was motivated (at least partly) by the epidemiological research of his time in developing his methods, and whether the latter influenced some working scientist.

In light of our discussion of SMD, one can even speculate that Mill's methods, and in particular his method of difference, might have provided the formal reconstruction of the basic "concepts and methods" of epidemiology that, according to Morabia, was missing. Indeed, Morabia acknowledges, without pursuing the issue, that "from Mill on, group comparisons combined with population thinking became a philosophically valid principle of knowledge acquisition" (2004, p. 102). Some encouraging suggestions in this direction are indeed to be found in the recent literature. For instance, Scholl (2013, p. 73) convincingly argues that Semmelweis, in his famous investigation of puerperal fever carried out in the years 1846–49, "made conscious use of something like the method of difference, and that he did so in full awareness of the method's logic and force".[7] Scholl provides a detailed reconstruction of Semmelweis' reasoning, based on the tables of mortality and morbidity appearing in his original publications, and concludes that "Semmelweis's data and arguments align neatly with [Mill's] methods" (2013, p. 67) and, in particular, that "the causal role of cadaverous matter was established according to the method of difference" (2013, p. 74). Similarly, Lee (2012) discusses Snow's study of the 1854 cholera outbreak in London (based on the so called Broad Street pump experiment), concluding that "Snow's experiment, as well as drug trials conducted today, appears to rely, in the main, on the method of difference" (2012, p. 153). A consonant assessment appears also in Tulodziecki's (2012) detailed discussion of the principles of reasoning underlying Snow's method.

[7] Semmelweis' discovery has been a favorite case study of philosophers of science, from Hempel (1966) to Gillies (2005) and beyond. Tulodziecki (2013) offers a very useful critical survey of all the most important philosophical reconstructions of this historical case. Both Salmon (1984, sec. 29) and Lipton (2004, pp. 74 ff.) discuss Semmelweis' method in relation with Mill's method of difference.

As the above cases show, Mill's method of difference is close to the kind of reasoning routinely implemented in science since the earliest days of epidemiology and statistics.[8] Indeed, it is not difficult to find references to Mill's methods also in later epidemiological writings (cf. Morabia 2004, p. 101, who refers in particular to the work of Mervyn Susser; see also Broadbent 2013, p. 67). To mention but one example, epidemiologist Raj Bhopal (2002, p. 117) recently claimed that "Mill's canons [...] are of paramount importance to epidemiology and are essentially incorporated into its own widely used criteria" and that, in particular, "the method of difference is at the core of epidemiological thinking (e.g. why do some people get heart disease and others of the same age and sex do not)". As Morabia (2004, p. 101) puts it, the method of difference is "typically the rationale for epidemiologic designs, either cohort or case-control studies, aiming to compare like with like. In epidemiology, however, the method has to be reformulated in probabilistic terms" (see also Morabia 2013). Our presentation of SMD provides precisely this kind of reformulation of Mill's MD.

*Early Insights on a Statistical Version of Mill's Method of Difference* As we already noted in Sect. 9.3, several logicians and philosophers have recognized that a probabilistic version of Mill's methods may illuminate some basic forms of inference used in scientific research. For instance, in a well-known textbook on critical thinking the author plainly claims, without further comments, that "the method of difference is virtually identical to the method of controlled experiment employed in such fields as biology, pharmacology, and psychology" (Hurley 2012, Ch. 10. pp. 540–41). Indeed, this seems to have become a commonplace in philosophical literature on evidence-based medicine, policy, and social science of the last few years. As an example, Cartwright and Hardie (2012, 33–34), introducing the reader of their *Evidence-based policy* to the very concept of an RCT, note that "an RCT is a study design based on John Stuart Mill's method of difference for making causal inferences", specifying that "an RCT is a Mill's method-of-difference group study in which individual units, all of which are supposed to be governed by the same causal principle, are randomly assigned to the treatment and control groups". Following them, Holman (2015, p. 1334), aiming at a "methodological definition of a placebo [...] determined by the underlying logic of ideal RCTs", plainly states that "the ideal RCT is a manifestation of Mill's method of difference". Similarly, in his introduction to the philosophy of economics, Reiss (2013, p. 201) explains that "RCTs aim to implement a probabilistic version of Mill's method of difference".

[8] Statisticians are familiar with Mill's method of difference at least since Sir Ronald Fisher (1890–1962) famously attacked it in his groundbreaking work on the *Design of experiment*s (1935). In his analysis of statistical causality, Holland (1986) discusses both Fisher and Mill as two of the central contributors to the topic. Rosenbaum (2009, Ch. 15) provides an assessment of the debate. For a compact but highly informative survey on the origins of control groups and RCTs, see Dehue (2005).

In view of the fact that most commentators appear to agree that the logic of RCTs and controlled inquiry is essentially Millian, it is surprising that, to the best of our knowledge, no previous attempts to formulate detailed statistical versions of the method of difference, or of the other Millian methods of eliminative induction, have been made. To be sure, one can find in the philosophical literature a couple of interesting suggestions on the need to integrate those methods with corresponding statistical versions. In his *Teatrise on probability* (1921), for instance, John Maynard Keynes claimed that an adequate inductive logic should be based on the theory of probability, since most "laws" established in scientific research are not universal but only statistical, i.e., what Keynes called "inductive correlations". As Keynes notes, his inductive correlations are the same as the "approximate generalizations", propositions of the form "Most *A* are *B*" discussed by Mill in his book (ch. XXIII). While recognizing Mill as a leading figure in inductive logic, Keynes also suggests that his ignorance of probability theory may have prevented him to develop an adequate theory of the inductive inference of approximate generalizations (1921, pp. 267–8).[9]

Von Wright's *Treatise on Induction and Probability* (1951) is probably the most systematic attempt to provide a logical reconstruction of Mill's methods, although perhaps not the clearest one (cf. Mackie 1980, p. 298; van Heuveln 2000, p. 28). In passing, von Wright (1951, pp. 183 ff.) notes that it would be interesting to investigate whether the "logic of induction" underlying Mill's methods could be applied to the assessment of "statistical inductions" or "statistical generalizations", his labels for Mill's approximate generalizations and Keynes' inductive correlations. However, he drops this suggestion without developing it.

Later, Salmon (1984, pp. 103 ff.) argued that, in order to account for the fundamental role played by statistical considerations within investigations concerning causal relations, one needs to put forward appropriate refined versions of Mill's methods. According to Salmon, one such refinement is the method of controlled experiment, whose hardcore he characterized as follows: if the presence of a certain condition *C* makes a difference for the onset of *E*—for instance, if *C* increases the probability that *E* is present—then *C* is a cause of *E*. His illustrative example, involving vitamin C and the common cold, is essentially identical to the one used to illustrate SMD in Sect. 9.3.3. Following Salmon, Grünbaum discusses Mill's methods in his analysis of inference patterns in medical and psychological sciences, noting that "the establishment of a causal connection in psychoanalysis, no less than in 'academic psychology' or in medicine, has to rely on modes of inquiry that

[9]Keynes' harsh assessment of Mill's contribution is worth quoting at length. After having said that "Mill has not justly apprehended the relativity of all inductive arguments to the evidence, nor the element of uncertainty which is present, more or less, in all the generalisations which they support", Keynes adds in footnote: "This misapprehension may be connected with Mill's complete failure to grasp with any kind of thoroughness the nature and importance of the theory of probability. The treatment of this topic in the *System of Logic* is exceedingly bad. His understanding of the subject was, indeed, markedly inferior to the best thought of his own time" (Keynes 1921, Pt. III, Ch. 23, pp. 267–68).

were refined from time-honored canons of causal inference pioneered by Francis Bacon and John Stuart Mill" (1984, p. 47). His conclusions are as clear as telling: "modernized or refined statistical versions of the famous four methods of controlled inquiry articulated by Mill serve to test the causal relevance of an *X* to a *Y*: such controlled inquiry shows whether the presence of X *makes a difference* to the occurrence of Y" (Grünbaum 1993, p. 163).

*The Cognitive Structure of Scientific Commonsense* There is a simple way to explain the wide, if often implicit, consensus on the plausibility of Mill's method of difference amongst epidemiologists and philosophers of different schools and tendencies. The reason, we suggest, is that Mill's method illuminates some of the essential aspects of what, following Kuipers (2001, p. x), we may call the cognitive structures of scientific inquiry. In different terms, the method of difference is a particularly clear illustration of what Kuipers (2000), Worrall (2007a) and others call "scientific commonsense", that is, a set of qualitative principles underlying such diverse theories of scientific method as Popperian falsificationism and Bayesianism, as well as the best scientific practice. According to Worrall one such principle, "an obvious (but in practice immensely powerful)" one, is the "idea that really telling evidence for any claim is evidence that at the same time tells against plausible rival alternative hypotheses" (Worrall 2007a, p. 1015; cf. also Howick 2011, Ch. 4). We can reformulate in a more precise way this fundamental methodological principle, widely embraced by scientists and philosophers of science, as follows: a hypothesis is strongly confirmed by empirical evidence if, and only if, the evidence is explained very well by the hypothesis, and it is not explained well by any plausible alternative hypothesis. This idea might be labeled as the "principle of explanatory differential".

It should be clear that the principle of explanatory differential underlies the idea of any controlled (clinical) study, and RCTs in particular (cf. Table 9.5 in Sect. 9.3.3). In this regard, Worrall (2007a, pp. 1015–1016) claims:

> In the case of therapeutic claims in medicine, this means evidence that tells in favour of the claim that the treatment at issue (or rather the 'characteristic features' of that treatment over and above any placebo effect) is responsible for the observed outcome, but that at the same time tells against plausible rival explanations of that observed outcome. [ . . . ] If the control group received no treatment at all (making it a 'natural history' group) then – depending, I would suggest, on the nature and size of the observed difference in average outcome – the alternative hypothesis may be plausible that the difference is a placebo effect; that is why clinical trials invariably involve a control group that is given either placebo or conventional treatment.

As readers will have noticed, similar considerations apply, more generally, to all inductive strategies for the testing of hypotheses of statistical causality as discussed in this paper. In fact, the principle of explanatory differential underlies both Mill's original method of difference MD (cf. also Howick 2008, Ch. 3) and its statistical version SMD. In the former case, the evidence provided by the relevant observation (cf. Table 9.1 in Sect. 9.3.1) rules out all other possible sufficient causes besides *C*, which is then inferred as the probable sufficient cause of phenomenon *E* of interest. In the latter case, things are only slightly more complicated. Indeed, what is ruled

out by applying SMD (cf. Table 9.3 in Sect. 9.3.2) is the hypothesis that the much greater frequency of $E$ in one sample with respect to the other is due, not to the causal factor $C$ under investigation, but to some other factors, for instance plain chance. In short, the conclusion that $C$ is a statistical cause of $E$ is made probable by excluding all other plausible alternative hypotheses concerning $E$ as improbable in the light of the relevant observation. In this way, the (statistical) method of difference illuminates a crucial aspect of the cognitive structures of scientific commonsense.

## 9.5 Concluding Remarks

In this paper, we offered a rational reconstruction of the logic of clinical trials and, more generally, of the assessment of hypotheses of statistical causality, along distinctively Millian lines. More precisely, we argued that an appropriate, statistical version of Mill's method of difference underlies most current methods used to test and evaluate efficacy hypotheses in many different fields, from pharmacology to evidence-based policy and social sciences. Very roughly, Mill's method of difference (MD) aims at identifying the sufficient cause of a given phenomenon by eliminating all other possible causes. Similarly, the statistical version of the method of difference (SMD) aims at identifying the cause responsible for a significant statistical difference between two samples from a given population by ruling out the alternative hypothesis that such difference is the result of chance. As we argued, this is the crucial insight underlying RCTs and many other kinds of clinical trials and, in general, methods of controlled inquiry widely adopted in different fields; indeed, SMD highlights a basic aspect of scientific commonsense or, in other terms, of the qualitative cognitive structures of scientific inquiry.

The reader may well wonder whether and how, as a matter of fact, Mill's work exerted a more or less direct influence on the historical trajectory that led to the development of randomized clinical trials. The brief historical overview offered in Sect. 9.4 shows that SMD-like intuitions have been around for quite a while in both philosophical and scientific literature—and this, we submit, renders our rational reconstruction all the more plausible. However, here we choose to remain noncommittal about the specifics of Mill's historical role.

We wish to conclude by mentioning that the foregoing discussion of SMD is just a preliminary contribution to a more extensive exploration of the basic logic of controlled inquiry, and that much work lies ahead in this regard, as the following examples suggest. First, in its present formulation SMD is just a sketch of a proper inference method, a sort of "inductive inference schema" which one needs to reconstruct in detail to put it at work. A more detailed discussion than the one offered in Sect. 9.3.2 is needed in order to assess how SMD is applied within the different statistical methods on the market and in the various kinds of test of statistical causality. Secondly, it is a task for the future to explore and define statistical versions of the other methods put forward by Mill: the "method of agreement" and that "of concomitant variations" look particularly promising under this respect. Finally, in

view of the consensus on the plausibility of Mill's insights in the field of hypothesis testing and causality assessment, it would be interesting to investigate particular examples where those insights are effectively applied by working scientists. For the moment, however, we remain content with having highlighted the Millian logic of clinical trials and controlled inquiry in general.

**Acknowledgements** We thank Theo Kuipers and four anonymous reviewers for very useful comments on a draft of this paper. Financial support from the Italian Ministry of Education, University and Research (R.F. and L.T.: PRIN grant "Models and Inferences in Science" no. 20122T3PTZ; G.C.: FIRB project "Structures and dynamics of knowledge and cognition" no. Turin unit: D11J12000470001; FFABR 2017 individual grant), and from the University of Turin and the Compagnia San Paolo (G.C.: project grant "Assessing information models: exploring theories and applications of optimal information search", no. D16D15000190005) is gratefully acknowledged.

## References

Benedetti, F. (2008). *Placebo effects. Understanding the mechanisms in health and disease*. Oxford: Oxford University Press.

Bhopal, R. J. (2002). *Concepts of epidemiology*. Oxford: Oxford University Press.

Bluhm, R. (2007). Clinical trials as Nomological machines: Implications for evidence-based medicine. In H. Kincaid & J. McKitrick (Eds.), *Establishing medical reality. Essays in the metaphysics and epistemology of biomedical science* (pp. 149–166). Dordrecht: Springer.

Broadbent, A. (2013). *Philosophy of epidemiology*. London: Palgrave Macmillan.

Brössel, P. (2013). The problem of measure sensitivity redux. *Philosophy of Science, 80*, 378–397.

Cartwright, N. (2007). Are RCTs the gold standard? *BioSocieties, 2*(1), 11–20.

Cartwright, N. (2012). RCTs, evidence, and predicting policy effectiveness. In S. Chow & J. Liu (Eds.), *Design and analysis of clinical trials. Concepts and methodologies*. Hoboken: Wiley.

Cartwright, N., & Hardie, J. (2012). *Evidence-based policy. A practical guide to doing it better*. Oxford: Oxford University Press.

Chow, S., & Liu, J. (2004). Design and analysis of clinical trials. In *Concepts and methodologies*. Hoboken: Wiley.

Crupi, V. 2016. Confirmation, *The Stanford encyclopedia of philosophy* (Fall 2016 ed.), Edward N. Zalta (Ed.), URL = http://plato.stanford.edu/archives/fall2016/entries/confirmation/

Day, S. (2007). *Dictionary for clinical trials* (2nd ed.). Chichester: Wiley.

Dehue, T. (2005). History of the control group. In *Encyclopedia of statistics in behavioral science*. Chichester: Wiley.

Festa, R., & Cevolani, G. (2017). Unfolding the grammar of Bayesian confirmation: Likelihood and Antilikelihood principles. *Philosophy of Science, 84*(1), 56–81.

Fitelson, B. (1999). The plurality of Bayesian measures of confirmation and the problem of measure sensitivity. *Philosophy of Science, 66*, S362–S378.

Fitelson, B., & Hitchcock, C. (2011). Probabilistic measures of causal strength. In P. M. Illari, F. Russo, & J. Williamson (Eds.), *Causality in the sciences* (pp. 600–627). Oxford: Oxford University Press.

Giere, R. N. (1979). Understanding scientific reasoning. In *Holt*. New York: Rinehart and Wiston.

Gillies, D. (2005). Hempelian and Kuhnian approaches in the philosophy of medicine: The Semmelweis case. *Studies in History and Philosophy of Biological and Biomedical Sciences, 36*(1), 159–181.

Grünbaum, A. (1993). *Validation in the clinical theory of psychoanalysis: A study in the philosophy of psychoanalysis*. Madison: International Universities Press.
Hempel, C. G. 1960. Inductive inconsistencies. *Synthese*, *12*, 439–469. Reprinted in *Aspects of scientific explanation*, Carl G. Hempel (pp. 53–79). New York/London: The Free Press/Collier Macmillan, 1965.
Hempel, C. G. (1966). *Philosophy of natural science*. Englewood Cliffs: Prentice-Hall.
Hilpinen, R. (1968). *Rules of acceptance and inductive logic*. Amsterdam: North-Holland Pub..
Holland, P. W. (1986). Statistics and causal inference. *Journal of the American Statistical Association, 81*(396), 945–960.
Holman, B. (2015). Why most sugar pills are not placebos. *Philosophy of Science, 82*, 1330–1343.
Howick, J. (2008). *Philosophical issues in evidence-based medicine: Evaluating the epistemological role of double blinding and placebo controls*. PhD thesis. London: London School of Economics.
Howick, J. (2011). *The philosophy of evidence-based medicine*. Wiley-Blackwell.
Howson, C., & Urbach, P. (2006). *Scientific reasoning. The Bayesian approach* (3rd ed.). La Salle: Open Court.
Hurley, P. J. (2012). *A concise introduction to logic* (11th ed.). Boston: Wadsworth Pub.
Jenkins, S. (2004). *How science works. Evaluating evidence in biology and medicine*. Oxford: Oxford University Press.
Kadane, J. B. (1996). *Bayesian methods and ethics in a clinical trial design*. New York: Wiley.
Kaptchuk, T. (1998). Powerful placebo: The dark side of the randomised controlled trial. *Lancet, 351*, 1722–1725.
Kaptchuk, T., et al. (2000). Do medical devices have enhanced placebo effects? *Journal of Clinical Epidemiology, 53*, 786–792.
Keynes, J. M. (1921). *A treatise on probability*. London: Macmillan & Co..
Kuipers, T. A. F. (2000). *From instrumentalism to constructive realism*. Dordrecht: Springer.
Kuipers, T. A. F. (2001). *Structures in science*. Dordrecht: Springer.
Lee, K. (2012). *The philosophical foundations of modern medicine*. London: Palgrave Macmillan.
Levi, I. (1967). *Gambling with truth.* Cambridge: MIT Press.
Lipton, P. (2004). *Inference to the best explanation* (2nd ed.). London/New York: Routledge/Taylor and Francis Group.
Macedo, A., et al. (2003). Placebo effect and placebos: What are we talking about? *European Journal of Clinical Pharmacology, 59*, 337–342.
Mackie, J. L. (1980). *The cement of the universe*. Oxford: Clarendon Press.
Mill, J. S. (1843). *A system of logic, ratiocinative and inductive* (8th ed.). London: Longmans, Green, and Company. 1886.
Miller, F., et al. (Eds.). (2013). *The placebo: A reader*. Baltimore: The Johns Hopkins University Press.
Morabia, A. (2004). Epidemiology: An epistemological perspective. In A. Morabia (Ed.), *A history of epidemiologic methods and concepts* (pp. 1–125). Basel: Birkhäuser Basel.
Morabia, A. (2013). Hume, Mill, Hill, and the sui generis epidemiologic approach to causal inference. *American Journal of Epidemiology, 178*(10), 1526–1532.
Reiss, J. (2013). *Philosophy of economics: A contemporary introduction*. New York: Routledge.
Romeijn, J.-W. (2017) Philosophy of statistics, *The Stanford encyclopedia of philosophy* (Spring 2017 ed.), Edward N. Zalta (Ed.), URL = https://plato.stanford.edu/archives/spr2017/entries/statistics/. Accessed on 3 Mar 2018.
Rosenbaum, P. R. (2009). *Design of observational studies*. Springer.
Rothwell, P. M. (2005). External validity of randomised controlled trials: To whom do the results of this trial apply? *Lancet, 365*, 82–93.
Rothwell, P. M. (2008). Factors that can affect the external validity of randomised controlled trials. *PLoS Clinical Trials*, *2006*.
Saint-Mont, U. (2015). Randomization does not help much, comparability does. *PLoS One, 10*(7), e0132102. https://doi.org/10.1371/journal.pone.0132102.
Salmon, W. C. (1984). *Logic* (3rd ed.). Englewood Cliffs: Prentice-Hall.

Scholl, R. (2013). Causal inference, mechanisms, and the semmelweis case. *Studies in History and Philosophy of Science, 44*, 66–76.

Skyrms, B. (2000). *Choice and chance: An introduction to inductive logic* (4th ed.). Scarborough: Wadsworth.

Sprenger, J. (2018). Foundations for a probabilistic theory of causal strength. *Philosophical Review, 127*(3), 371–398.

Teira, D. (2011). Frequentist versus Bayesian clinical trials. In F. Gifford (Ed.), *Handbook of the philosophy of science* (Philosophy of medicine) (Vol. 16, pp. 255–297). Amsterdam: Elsevier.

Tulodziecki, D. (2012). Principles of reasoning in historical epidemiology. *Journal of Evaluation in Clinical Practice, 18*(5), 968–973.

Tulodziecki, D. (2013). Shattering the myth of Semmelweis. *Philosophy of Science, 80*(5), 1065–1075.

Urbach, P. (1985). Randomization and the design of experiments. *Philosophy of Science, 52*, 256–273.

van Heuveln, B. (2000). A preferred treatment of Mill's methods: Some misinterpretations by modern textbooks. *Informal Logic, 20*, 19–42.

von Wright, G. H. (1951). *A treatise on induction and probability*. London: Routlegde & Kegan Paul.

Worrall, J. (2007a). Evidence in medicine and evidence-based medicine. *Philosophy Compass, 2*, 981–1022.

Worrall, J. (2007b). Why there's no cause to randomize. *The British Journal for the Philosophy of Science, 58*, 451–488.