Bayesian Expectancy Invalidates Double-Blind Randomized Controlled Medical Trials*

Gilles Chemla
Imperial College Business School, DRM/CNRS, and CEPR.

Christopher A. Hennessy
London Business School, CEPR, and ECGI.

June 2016

Abstract

Double-blind RCTs are viewed as the gold standard in eliminating placebo effects and identifying non-placebo physiological effects. Expectancy theory posits that subjects have better present health in response to better expected future health. We show that if subjects Bayesian update about efficacy based upon physiological responses during a single-stage RCT, expected placebo effects are generally unequal across treatment and control groups. Thus, the difference between mean health across treatment and control groups is a biased estimator of the mean non-placebo physiological effect. RCTs featuring low treatment probabilities are robust: Bias approaches zero as the treated group measure approaches zero.

*We thank seminar participants at UCLA and Zurich, John Rust, Jacob Sagi, and Jan Starmans for feedback. Thanks to Bruce Carlin for a medical doctor/economist perspective. Corresponding author (Hennessy): Regent's Park, London, NW1 4SA, U.K.; chennessy@london.edu; 44(0)2070008285. This research was supported in part by a European Research Council Grant (Hennessy).

1 Introduction

A critical stated objective for medical researchers is to measure the non-placebo physiological effect of a treatment, also known in the medical literature as *characteristic effect* (Grünbaum (1986)) or *specific effect* (Malani (2006)). Since Fisher (1935), the double-blind randomized controlled trial (RCT below) has been viewed as the gold standard in estimating non-placebo physiological effects. In fact, in describing the rise of RCTs in medicine in *The Lancet*, Kaptchuk (1998) writes, "The greater the placebo's power, the more necessity there was for the masked RCT itself." In the U.S., E.U. and Japan, the gold standard status of RCTs is codified under the International Conference on Harmonization of Technical Requirements for Registration of Pharmaceuticals for Human Use (ICH, 2000). The ICH writes, "Control groups have one major purpose: to allow discrimination of patient outcomes [...] caused by the test treatment from outcomes caused by other factors, such as the natural progression of the disease, observer or patient expectations, or other treatment." In fact, the perceived reliability of the RCT has caused the methodology to be emulated in other disciplines. For example, the RCT is held up by Angrist and Pischke (2009) as the ideal for achieving unbiased estimates of causal effects in the social sciences.

The logical argument for the double-blind medical RCT is by now so familiar that it typically escapes discussion. Health quality is viewed as the sum of the non-placebo physiological effect plus any brain-modulated physiological ("mental" or "placebo") effect.[1] If subjects are randomly assigned across treatment and control groups, and blind to their actual assignment, then the expectation of mental effects is posited to be equal across treatment and control groups. It would follow then that the difference between mean health outcomes across groups yields an unbiased estimate of the expectation of the non-placebo physiological effect.

The traditional proof of RCT validity assumes mental effects are additive identically distributed random variables independent of the assigned group. However, as shown below, the critical indepen-

[1] Following the literature, we use these three terms interchangeably noting that placebo effects should not be confused with inert pills (placebos).

dence assumption is at odds with a rational Bayesian formulation of *expectancy theory*, the leading theory of placebo effects. Stewart-Williams and Podd (2004) state, "On the expectancy account, the effects of such factors come through their influence on the placebo recipient's expectancies." Malani (2006) writes, "placebo effects cause more optimistic patients to respond better to treatment than less optimistic patients." Similarly, in perhaps the first known placebo-controlled trial, Haygarth (1801) wrote that his study, "clearly prove[d] what wonderful effects the passions of hope and faith, excited by mere imagination, can produce on disease."

The type of beneficial brain-modulated physiological effects posited by Haygarth (1801) are often treated as irrational noise terms. Instead, we here consider that such effects arise rationally from the expectation of better future health quality arising from improvements in the efficacy of future treatments. For example, expectation of less future pain may reduce stress, improving outcomes today for subjects suffering from ulcers or hypercholesterolemia. Similarly, expectation of higher future survival rates may alleviate the severe anxiety associated with life-threatening diseases, with relaxation, rest and sleep improving health outcomes today. As a final example, expectation of improved future brain functioning, that is, alleviation of hopelessness, may mitigate depression today.

We show that the expectation of mental effects should not be presumed to be equal across treatment and control groups in RCTs. Intuitively, with rational Bayesian test subjects, beliefs about a treatment's efficacy will vary systematically with the probability distribution governing the respective physiological states induced by the treatment and control medications.[2] Unless the probability distributions are equal across treatment and control groups, beliefs will differ across them, implying expected mental effects are not equal. The difference in average health outcomes across treatment and control groups then delivers a biased estimate of the mean of the non-placebo physiological effect.

There have been few attempts by economists to formally model placebo effects. An important exception is Malani (2006). However, his objective was to develop a formal empirical test for

[2]Similar biases emerge under biased updating, e.g. overconfidence, but we here stress rationality.

placebo effects. The key prediction he derives from his model is that, if there are indeed placebo effects, then treatment and control groups should have better outcomes in trials featuring a higher treatment probability. After confirming this prediction for the treatment group, Malani states, "An important footnote to the findings reported in this paper is that regression analysis also reveals a positive correlation between the probability of treatment and outcomes in the control group, but only when the control therapy is active treatment, namely H_2 blockers in PPI trials. It does not find a positive correlation when the control is an inert pill." That is, Malani's empirical findings suggest that placebo effects differ across treatment and control groups in a manner dependent upon the true non-placebo physiological effect of the control. As shown below, such cross-group placebo differentials are inconsistent with standard additive placebo effects, such as those modeled by Malani (2006). In this paper, we offer a rational Bayesian theory for such differentials, one that calls into question the logical basis for the presumption of lack of bias in RCTs.

There is a large medical literature on RCTs. Rothwell (2005, 2006) provides excellent surveys. Existing critical examinations of medical RCTs have emphasized the difficulties in their practical implementation. For example, Rothwell discusses numerous practical sources of selection bias, such as attrition of subjects in multi-stage trials. Further, Rothwell also discusses the difficulty in measuring comprehensive health outcomes accurately. More formally, Chan and Hamilton (2006) develop a dynamic discrete choice model to account for the effects of selection bias in multi-stage trials, as well as incorporating the possibility of unobservable side effects relevant to a true measure of health quality. Such criticisms and analyses apply to the difficulty in implementing RCTs but they do not directly challenge the logical basis of the RCT methodology itself. For clarity, our model is constructed to rule out selection bias and attrition, and assumes health quality is measurable.

Consistent with our model, there is a large empirical medical literature documenting discrepancies between RCT outcomes and actual long-term outcomes in relation to rheumatoid arthritis, osteoporosis, post-myocardial infarction, multiple sclerosis, and heart failure (Pincus (1998), Rothwell (2005)). Our results show such empirical regularities may not be due to imperfect RCT

implementation, but rather may point to a more fundamental methodological problem.

Finally, our paper contributes at the margin to the recent literature on optimal experiment design, as surveyed and extended by Chassang, Padro i Miguel, and Snowberg (2015), who consider endogenous effort and hidden information in RCTs using a principal-agent approach. As we show, RCTs are subject to bias due to Bayesian updating by agents. However, we show this bias goes to zero as the measure of the treatment group goes to zero. Intuitively, if the probability of being in the treatment group is infinitesimal, then beliefs regarding the new drug's efficacy are insensitive to a subject's realized health quality during the RCT, so the statistical distribution of mental effects will be nearly identical across treatment and control groups. Our result thus highlights a potentially overlooked cost associated with the common practice of adopting balanced panels, or in using a higher treatment probability to attract voluntary participation.

2 The Model

All aspects of the experimental setting are common knowledge.[3] There are two dates $d \in \{1, 2\}$. At $d = 1$, a double-blind randomized, parallel group, controlled trial (RCT) is conducted. The objective of the RCT is to measure the efficacy of a novel drug.[4] The measured efficacy determines the probability of the tested drug being distributed at $d = 2$, as well as the probability of next-generation improvements to the drug. Depending on the setting, one can think of the control group as being given either an inert drug (placebo-controlled trial) or some traditionally-used drug (controlled trial).

In the model, the RCT is ideal, with the test panel being equivalent to the afflicted population, eliminating self-selection concerns.[5] The afflicted population $\mathcal{I}$ is a measure one continuum of ex ante identical agents, eliminating potential concern over small sample bias.[6] Below, $\mathcal{T}$ ($\mathcal{C}$) denotes

[3] This is consistent with commonly-imposed informed consent laws.

[4] We consider drugs to fix ideas, but the analysis applies to medical treatments generally.

[5] See Malani (2006) for a detailed analysis of self-selection in RCTs.

[6] Deaton (2010) expresses concern over small sample biases in RCTs.

the set of agents randomly assigned to the treatment (control) group. The measure of the treatment group is $t \in (0,1)$.

Just after taking her assigned drug, agent $i \in \mathcal{I}$ experiences her respective *direct physiological state* p_i^1, a random variable with support $\mathcal{P} \equiv \left[\underline{p}, \overline{p}\right]$. The direct physiological state represents the health quality that would be experienced by the agent in the absence of any mental effect. If $i \in \mathcal{C}$, then p_i^1 is an independent draw from an atomless twice continuously differentiable cumulative distribution F_C, with probability density f_C. If $i \in \mathcal{T}$, then p_i^1 is an independent draw from an atomless twice continuously differentiable cumulative distribution F_S, with probability density f_S. Here S denotes the *efficacy state* of the new drug. The efficacy state is not known at the start of the RCT. It is common knowledge that $S \in \{L, H\}$, with L (H) denoting low (high) efficacy. Let $\lambda \equiv \Pr[S = H] = \lambda$, and assume $\lambda \in (0,1)$.

Letting

$$\mu_J \equiv \int_{\mathcal{P}} p f_J(p) dp,\ J \in \{C, L, H\},$$

we assume

$$\mu_H \geq \mu_L.$$

The following technical assumption is adopted.

Assumption 1: *For each $p \in \mathcal{P}$ there exists a $J \in \{C, L, H\}$ such that $f_J(p) > 0$. For each $p \in (\underline{p}, \overline{p})$, $f_J(p) > 0$ for all $J \in \{C, L, H\}$.*

The first part of Assumption 1 ensures beliefs are well-defined on $\mathcal{P}$. The second part ensures the derivative of beliefs is well-defined on $(\underline{p}, \overline{p})$.

Health quality is measured without error. During the RCT ($d = 1$) agent $i \in \mathcal{I}$ has health quality Q_i^1, where

$$Q_i^1 \equiv p_i^1 + M_i. \tag{1}$$

In the preceding equation, the first term, the direct physiological state, is assumed to be privately observed by the agent. The second term captures the mental effect. The assumption that the agent observes p_i^1 is not necessary if beliefs, to be discussed, are monotone in p_i^1. With monotone beliefs,

the agent can invert Q_i^1 and infer her direct physiological state. Since additivity of the mental effect plays an important role in the traditional proof for the unbiasedness of RCTs, we now stress this assumption.

Assumption 2: *The brain-modulated physiological effect (mental effect) enters health quality additively.*

Although not our focus, we note the additivity assumption may be questionable in some medical contexts, as noted by Malani (2006) and Chassang, Snowberg, Seymour and Bowles (2015). For example, in settings with upper (complete recovery) and lower (death) bounds on the health quality variable, max and min functions would be applied to $p + M$. Similarly, one might argue a multiplicative functional form, pM, is appropriate in some clinical settings. Under such functional form assumptions, the expectation of mental effects would not be equal across treatment and control groups even if the M_i were, in fact, i.i.d. random variables.

For simplicity, events at $d = 2$ are modeled in reduced-form. If the efficacy state is S, then with probability π_S, next-generation improvements to the novel drug will be implemented and the newly-improved drug will be distributed at $d = 2$.[7] The benefit of improvements to the drug is to increase the health quality of those who take it by an increment δ_S, with $\delta_H \geq \delta_L \geq 0$. Continuation is more likely if the efficacy state is H, specifically $\pi_H \geq \pi_L \geq 0$. We assume $\mu_L + \delta_L \geq \mu_C$. Assuming, as we do, that drugs come at zero price to the consumer (due to, say, insurance) and that agents are risk neutral, the latter inequality ensures that the newly-improved drug will be consumed voluntarily if it is offered at $d = 2$.

For simplicity, assume the effect of both the treatment and control drug is non-cumulative. In particular, conditional upon the efficacy state S, for each agent $i \in \mathcal{I}$, and for each drug type, the direct physiological state experienced at $d = 2$ is independent of the direct physiological state experienced at $d = 1$. That is, p_i^2 is an independent draw of the direct physiological state from the relevant distribution, specifically F_S for the new drug and F_C for the control.

[7] One could allow the probability of distributing the drug to differ from the probability of improvements. One could also allow for future improvements to the control. This would only complicate the algebra and add notation.

Let Φ_S be an indicator for continuation with the novel drug (next-generation improvements plus distribution at $d = 2$). We then have:

$$\pi_S \equiv \Pr[\Phi_S = 1].$$

Health quality at $d = 2$ can then be expressed as:

$$Q_i^2 \equiv p_i^2 + \Phi_S \delta_S. \tag{2}$$

Let β denote the probability assessment of an agent that the efficacy state is H based upon the direct physiological state experienced by this same agent during the RCT. From Bayes' rule we have:

$$\beta(p) \equiv \Pr[S = H | p_i^1 = p] = \frac{\lambda \left[t f_H(p) + (1-t) f_C(p) \right]}{t \left[\lambda f_H(p) + (1-\lambda) f_L(p) \right] + (1-t) f_C(p)}. \tag{3}$$

For brevity, let:

$$X(p) \equiv \mathbb{E}[Q_i^2 | p_i^1 = p].$$

Given the stated assumptions, we have:

$$X(p) = \beta(p)[\pi_H(\mu_H + \delta_H) + (1 - \pi_H)\mu_C] + [1 - \beta(p)][\pi_L(\mu_L + \delta_L) + (1 - \pi_L)\mu_C]. \tag{4}$$

Two points are worth noting in the preceding equation. First, expected future health quality varies with the continuation probabilities (π_L, π_H). Second, expected future health capitalizes not only the effects of the new drug but also anticipated improvements to this drug, as captured by the improvement parameters (δ_L, δ_H).

The following assumption describes the mapping from expected future health quality to the present-day mental effect.

Assumption 3: *Brain-modulated physiological effects (mental effects) are equal to* $\Psi(X)$, *with* Ψ *being continuously differentiable and strictly increasing.*

Under Assumption 3, the mental component of health quality at $d = 1$ can be computed as:

$$M(p) = \Psi \left[X(p) \right] \Rightarrow M_i = M(p_i^1). \tag{5}$$

Notice, as reflected in the preceding equation, in the present formalization of expectancy theory, beneficial brain-modulated physiological effects are driven by optimism about the efficacy of future medication, i.e. hope, not by beliefs regarding whether one is receiving the treatment or control during the present-day RCT. The key probability assessment is β, since this belief determines the expectation of Q_i^2.

The objective of the RCT is to estimate the expectation of the direct (non-placebo) effect of the drug on health quality. Under the stated assumptions we have:

$$\text{Direct (Non-Placebo) Effect} \equiv \mathbb{E}[p_i^1 | i \in \mathcal{T}, S] - \mathbb{E}[p_i^1 | i \in \mathcal{C}, S] = \mu_S - \mu_C. \tag{6}$$

The expected treatment-control health quality difference is:

$$\mathbb{E}[Q_i^1 | i \in \mathcal{T}, S] - \mathbb{E}[Q_i^1 | i \in \mathcal{C}, S] = \underbrace{\mu_S - \mu_C}_{\text{Direct Effect}} + \underbrace{\{\mathbb{E}[M_i | i \in \mathcal{T}, S] - \mathbb{E}[M_i | i \in \mathcal{C}, S]\}}_{\text{Bias}}, \tag{7}$$

where

$$\mathbb{E}[M_i | i \in \mathcal{T}, S] - \mathbb{E}[M_i | i \in \mathcal{C}, S] = \int_{\mathcal{P}} M(p)[f_S(p) - f_C(p)]dp. \tag{8}$$

It is traditional to think of mental effects, the M_i, as being i.i.d. random variables. Under the traditional interpretation, the bias term in equation (7) is equal to zero, implying the mean treatment-control health quality difference yields an unbiased estimate of the mean of the direct non-placebo physiological effect of the treatment relative to the control. Thus, the absence of bias in RCTs can be understood as being predicated upon two assumptions: additivity and i.i.d. mental effects.

In order to highlight the main causal mechanism, it is instructive to first derive a condition under which absence of bias is assured.

Proposition 1 *A sufficient condition for the expected treatment-control health quality difference to equal the expectation of the direct physiological effect, regardless of the true efficacy state S, is*

$$\pi_H(\mu_H + \delta_H) + (1 - \pi_H)\mu_C = \pi_L(\mu_L + \delta_L) + (1 - \pi_L)\mu_C.$$

Proof. $M'(p) = \Psi'[X(p)]X'(p)$. Under the stated condition $X' = 0$. So the bias term in equation (7) is 0. ■

The intuition for the preceding proposition is as follows. Under the stated condition, expected future health quality is equal across efficacy states H and L, implying differences between treatment and control group beliefs (β) regarding the efficacy state S are inconsequential.

In light of the preceding proposition, the remainder of the analysis imposes the following additional assumption.

Assumption 4: *The conditional expectation of future health quality is strictly higher in efficacy state H than in efficacy state L*:

$$\pi_H(\mu_H + \delta_H) + (1 - \pi_H)\mu_C > \pi_L(\mu_L + \delta_L) + (1 - \pi_L)\mu_C.$$

3 Bias in RCTs

We now describe a number of settings in which RCTs have a bias that is easy to sign. In each setting described below, the belief function β will be shown to be strictly increasing, implying the mental effect function M is also strictly increasing. That is, the technology is such that a better draw of p_i^1 during the RCT causes the agent to assign a higher probability to efficacy state H, which leads to better brain-modulated physiological responses. With M demonstrated to be increasing, we use stochastic dominance relationships across the distribution functions to determine the sign of the bias.

For brevity, let:

$$R_S(p) \equiv \frac{f_S(p)}{f_C(p)} \ \forall \ S \in \{L, H\} \text{ and } p \in (\underline{p}, \overline{p}).$$

Differentiating β and rearranging terms one finds:

$$\text{Sign}[\beta'(p)] = \text{Sign}\left[\frac{[tR_H(p) + (1-t)]'}{tR_H(p) + (1-t)} - \frac{[tR_L(p) + (1-t)]'}{tR_L(p) + (1-t)}\right]. \tag{9}$$

Using the preceding equation, we have the following proposition.

Proposition 2 *If* $\frac{f_H}{f_C}$ *and* $\frac{f_C}{f_L}$ *are strictly increasing (MLRP), then the expected treatment-control health quality difference is greater (less) than the expectation of the direct physiological effect in efficacy state* H (L)*.*

Proof. Under the stated conditions it follows from equation (9) that β is strictly increasing on $(\underline{p}, \overline{p})$, from which it follows M is also strictly increasing on $(\underline{p}, \overline{p})$. Since MLRP implies FOSD, the biases then follow from the fact that the distribution F_C first-order dominates F_L and is first-order dominated by F_H.■

The intuition for the preceding proposition is as follows. In efficacy state H (L), agents assigned to the treatment group draw their p_i^1 from a distribution that dominates (is dominated by) the distribution from which the control group draws. In expectation, this causes them to assess a higher (lower) probability of efficacy state H and to have more (less) favorable brain-modulated physiological responses.

The next proposition presents conditions under which there will be upward bias in both states.

Proposition 3 *If* $\frac{f_H}{f_L}$ *is strictly increasing (MLRP) and*

$$\left(\frac{f_H}{f_C}\right)'(p) > \left(\frac{f_L}{f_C}\right)'(p) > 0 \text{ for all } p \in (\underline{p}, \overline{p}),$$

then the expected treatment-control health quality difference is greater than the expectation of the direct physiological effect in efficacy state L *and efficacy state* H*.*

Proof. From FOSD it follows that in order to establish the entire result we must simply verify β is strictly increasing on $(\underline{p}, \overline{p})$. Rearranging terms in equation (9), we find that β is strictly increasing iff

$$\begin{aligned} 0 &< R'_H(p)\left[tR_L(p) + (1-t)\right] - R'_L(p)\left[tR_H(p) + (1-t)\right] \\ &\Leftrightarrow \quad 0 < t\left[R_L(p)R'_H(p) - R'_L(p)R_H(p)\right] + (1-t)\left[R'_H(p) - R'_L(p)\right] \\ &\Leftrightarrow \quad 0 < t[R_L(p)]^2\left[\frac{f_H(p)}{f_L(p)}\right]' + (1-t)\left[R'_H(p) - R'_L(p)\right]. \blacksquare \end{aligned}$$

The intuition for the preceding proposition is as follows. In both possible efficacy states, agents assigned to the treatment group draw their p_i^1 from a distribution that dominates the distribution from which the control group draws. In expectation, this causes them to assess a higher probability of efficacy state H and to have more favorable brain-modulated physiological responses.

Finally, we describe a case in which $\mu_L = \mu_C$, yet, under the stated assumptions, the expected treatment-control health quality difference is positive in state L. The construction is as follows. Suppose $\mathcal{P} = [0, 1]$, with $f_L(p) > f_C(p)$ for $p > 3/4$. Consider then Ψ having a convex kink at zero at the point $\beta(3/4)$. Under this type of convex brain-modulated physiological response, the treatment has a stronger mental effect than the control even in state L.

Remark 1 *Let* $\mathcal{P} = [0, 1]$, *with* $f_H = 2p$; $f_L = 1$; $f_C = 4p$ *for* $p \leq 1/2$; *and* $f_C = 4(1-p)$ *for* $p > 1/2$ *implying* $\mu_L = \mu_C$. *Suppose further*

$$\begin{aligned} \Psi(X) &\equiv \max\{X - X^*, 0\}, \\ X^* &\equiv \frac{\pi_L}{2} + \left[\frac{\lambda + \frac{\lambda t}{2}}{1 + \frac{\lambda t}{2}}\right]\left[\frac{2\pi_H}{3} - \frac{\pi_L}{2}\right]. \end{aligned}$$

Then the expectation of the treatment-control health quality difference is greater than the direct physiological effect in efficacy state L, *as well as efficacy state* H.

Proof. We first verify β is increasing, implying so too is M. Let

$$\begin{aligned} \gamma(p) &\equiv \frac{tR_H(p) + (1-t)}{tR_L(p) + (1-t)} \; \forall \; p \in (\underline{p}, \overline{p}) \\ &\Rightarrow \ln[\gamma(p)] = \ln[tR_H(p) + (1-t)] - \ln[tR_L(p) + (1-t)] \\ &\Rightarrow \frac{\gamma'(p)}{\gamma(p)} = \frac{[tR_H(p) + (1-t)]'}{[tR_H(p) + (1-t)]} - \frac{[tR_L(p) + (1-t)]'}{[tR_L(p) + (1-t)]}. \end{aligned}$$

From equation (9), it follows the sign of the slope of β is equal to the sign of the slope of γ. Consider first $p < 1/2$. We have the following increasing function:

$$\gamma(p) = \frac{tR_H(p) + (1-t)}{tR_L(p) + (1-t)} = \frac{1 - \frac{t}{2}}{\frac{t}{4p} + (1-t)}$$

Consider next $p > 1/2$. We have the following increasing function:

$$\gamma(p) = \frac{tR_H(p) + (1-t)}{tR_L(p) + (1-t)} = \frac{2tp + 4(1-t)(1-p)}{t + 4\,(1-t)\,(1-p)}.$$

The rest of the proof follows from the description preceding the remark.■

4 Bias and Treatment Group Measure

This section briefly examines the relationship between bias and the experiment design parameter t. Traditionally, the desire to maximize statistical power in finite samples has led to the adoption of treatment and control groups of equal size. Ethical concerns regarding leaving some subjects untreated in placebo-controlled trials has at times led to the adoption of $t > 1/2$. Finally, the desire to attract voluntary participation has also led researchers to utilize $t > 1/2$. However, the following proposition illustrates a benefit to trials featuring small t.

Proposition 4 *As the measure of the treatment group goes to zero, bias goes to zero.*

Proof. As t tends to zero, β tends to λ, and β' tends to zero. We recall then $M'(p) = \Psi'[X(p)]X'(p)$. Under the stated condition, X' tends to zero and so too does the bias term in equation (7).■

Intuitively, bias arises from unequal expected mental effects across treatment and control groups, with differences arising from differences in the distribution of Bayesian beliefs. However, with t close to zero, subjects place little weight on their own draw of p_i^1 in forming beliefs about the efficacy state S. With such an experiment design, the statistical distribution of beliefs and mental effects will be nearly identical across treatment and control groups, virtually eliminating bias.

Based on the preceding proposition, a natural question to ask is whether bias is increasing in t. To address this question, we now express the bias in state $S \in \{L, H\}$ as a function of the trial design parameter t:

$$B_S(t) \equiv \int_{\mathcal{P}} M(p,t)[f_S(p) - f_C(p)]dp. \tag{10}$$

Differentiating the preceding equation we obtain:

$$B'_S(t) = \left[\begin{array}{l} [\pi_H(\mu_H + \delta_H) + (1 - \pi_H)\mu_C] \\ -[\pi_L(\mu_L + \delta_L) + (1 - \pi_L)\mu_C] \end{array} \right] \int_{\mathcal{P}} \Psi' [X(p,t)] \frac{\lambda(1-\lambda) f_C(p)[f_H(p) - f_L(p)][f_S(p) - f_C(p)]}{[t(\lambda f_H(p) + (1-\lambda) f_L(p)) + (1-t) f_C(p)]^2} dp. \tag{11}$$

From the preceding expression, we have the following two propositions presenting sufficient conditions for the absolute value of bias to be increasing in t.

Proposition 5 *Suppose $\frac{f_H}{f_C}$ and $\frac{f_C}{f_L}$ are strictly increasing (MLRP), resulting in positive (negative) bias in efficacy state H (L). Then if the probability densities have a single crossing point p^* at which $f_H(p^*) = f_L(p^*) = f_C(p^*)$, the absolute value of bias in both states is strictly increasing in t.*

Proof. The result follows from equation (11) and the fact that under the stated assumptions:

$$\begin{aligned} p &< p^* \Rightarrow f_H(p) < f_C(p) < f_L(p) \\ p &> p^* \Rightarrow f_H(p) > f_C(p) > f_L(p). \blacksquare \end{aligned}$$

Proposition 6 *Suppose $\frac{f_H}{f_L}$ is strictly increasing (MLRP) and*

$$\left(\frac{f_H}{f_C}\right)'(p) > \left(\frac{f_L}{f_C}\right)'(p) > 0 \text{ for all } p \in (\underline{p}, \overline{p}),$$

so that bias in both efficacy states is positive. Then if the probability densities have a single crossing point p^* at which $f_H(p^*) = f_L(p^*) = f_C(p^*)$, the bias in both states is strictly increasing in t.

Proof. The result follows from equation (11) and the fact that under the stated assumptions:

$$\begin{aligned} p &< p^* \Rightarrow f_H(p) < f_L(p) < f_C(p) \\ p &> p^* \Rightarrow f_H(p) > f_L(p) > f_C(p). \blacksquare \end{aligned}$$

Notwithstanding the preceding two propositions, it is readily verified that the absolute value of bias is not necessarily increasing in the trial design parameter t. To see this, note that if the probability densities do not have a single crossing point, as the two propositions assume, then the term $(f_H - \partial f_L)(f_S - f_C)$ in the integrand in equation (11) is potentially negative on some intervals. By letting the slope of Ψ go to zero for p outside all such intervals one obtains $B'_S(t) < 0$.

5 Concluding Remarks

This paper illustrates a fragility associated with double-blind RCTs, often viewed as the gold standard in medicine for estimating pure non-placebo physiological effects (characteristic or specific effects). Specifically, when positive expectancy about future health quality leads to better present-day health quality, then the expectation of mental effects cannot be presumed equal across treatment and control groups in RCTs, since beliefs will vary systematically with the distribution of direct physiological states. It follows that the difference between mean health outcomes across treatment and control groups is a biased estimator of the mean of the direct (non-placebo) physiological effect.

Before closing, it is worth discussing why it would be, as a general matter, inappropriate to credit a studied drug with the mental effects measured during an RCT. First and foremost, regulators have stated that their goal is to strip out mental effects–perhaps due to concern over manipulability of emotional states. Second, as our analysis shows, the expectancy of medical subjects is related to their assessment of the probability of approval and production of a drug, captured by the model parameters (π_L, π_H). In reality, these parameters are likely to vary over time and cross-sectionally with the financial constraints of companies, regulatory stringency, and governmental funding capacity. They do not represent physiological constants. Third, as argued above, and as shown in equation (4), expectancy in a current RCT reflects in part the value of the control, as well as the value of the treatment in counter-factual states. Fourth, as was shown, expectancy during a current RCT reflects the anticipated value of next-generation drugs, not just the value of the drug being studied.

References

[1] Angrist, Joshua D. and Jorn-Steffen Pischke, 2009, *Mostly Harmless Econometrics: An Empiricist's Companion,* Princeton University Press.

[2] Chan, Tat Y. and Barton H. Hamilton, 2006, Learning, Private Information, and the Economic Evaluation of Randomized Experiments, *Journal of Political Economy* 114 (6), 997-1040.

[3] Chassang, Sylvain, Erik Snowberg, Ben Seymour, and Cayley Bowles, 2015, Accounting for Behavior in Treatment Effects: New Applications for Blind Trials, *PLoS ONE* 10 (6).

[4] Chassang, Sylvain, Gerard Padro i Miguel, and Erik Snowberg, 2015, Selective Trials: A Principal-Agent Approach to Randomized Controlled Experiments, *American Economic Review.*

[5] Deaton, Angus, 2010, Instruments, Randomization, and Learning about Development, *Journal of Economic Literature* 48 (2), 424-455.

[6] Fisher, Ronald A., 1935, *The Design of Experiments.* London: Oliver and Boyd.

[7] Grünbaum, Adolf, 1986, The Placebo Concept in Medicine and Psychiatry, *Psychological Medicine* 16 (1), 19-38.

[8] Haygarth, John, 1801, *Of the Imagination as a Cause and as a Cure of Disorders of the Body: Exemplified by Fictitious Tractors and Epidemical Convulsions.* Bath: Crutwell.

[9] International Conference of Harmonization, 2000, Choice of Control Group and Related Issues in Clinical Trials E10, Department of Health and Human Services: Center for Biological Evaluation and Research.

[10] Kaptchuk, Ted J., 1998, Powerful Placebo: The Dark Side of the Randomised Controlled Trial, *The Lancet* 351, 1722-1725.

[11] Malani, Anup, 2006, Identifying Placebo Effects with Data from Clinical Trials, *Journal of Political Economy* 114 (2), 236-256.

[12] Pincus, Theodore, 1998, Rheumatoid Arthritis: Disappointing Long-Term Outcomes Despite Successful Short-Term Clinical Trials, *Journal of Clinical Epidemiology* 41, 1037-1041.

[13] Rothwell P.M., 2005, External Validity of Randomised Controlled Trials: To Whom do the Results of this Trial Apply?, *The Lancet* 365, 82-95.

[14] Rothwell P.M., 2006, Factors That Can Affect the External Validity of Randomised Controlled Trials, *PLOS Clinical Trials*, 1-5.

[15] Stewart-Williams, Steve, and John Podd, 2004, The Placebo Effect: Dissolving the Expectancy versus Conditioning Debate, *Psychological Bulletin* 130 (2), 324-340.