

Patient enrollment in medical trials: Selection bias in a randomized experiment

Anup Malani*

University of Virginia and University of Chicago

February 8, 2006

Abstract. Self-selection can bias estimates of treatment effects from randomized experiments if one is interested not merely in the effect of treatment on the treated, but in extrapolating results to the general population. This paper employs the Roy model to study this problem in the context of medical trials. The main insight is that, as the probability of receiving active treatment rises, patients who are less optimistic about new treatment will begin to enroll and estimates of treatment effects will fall. This, in turn, implies that selection bias is positive. These findings are confirmed with data from trials of ulcer medications.

JEL Classification Numbers: I10, C90, C31.

Keywords: selection, randomization bias, experiment, treatment effect.

*1111 E. 60th Street, Chicago, IL 60605, (773) 702-9602 (phone), (773) 702-0730 (fax), amalani@virginia.edu.

Conventional wisdom holds that patient self-selection does not pose a problem for randomized medical trials. The reason is that random assignment eliminates unobservable differences between treatment groups. This is sufficient protection against bias if the researcher only cares about the effect of treatment on the treated – i.e., the effect of treatment on patients enrolled in the trial. Most trials, however, are concerned with predicting the effect of treatment on a patient population beyond the trial. In these cases, randomization does not prevent selection bias¹ which arises from systematic differences between the trial population and the general patient population.²

This paper proposes a model of patient self-selection into clinical trials in order to gain insight on the effect of this sorting on standard estimates of treatment effects. The model is based upon the Roy model of sorting of workers across sectors – with patients playing the role of workers, and treatment via a trial and treatment outside the trial as the two sectors. My analysis makes two critical assumptions. First, there is a new treatment that is being tested (against a placebo control) in a medical trial. The new treatment is only available through the trial. There may be a conventional treatment, however, that is available outside the trial. Second, the population of interest, i.e., the population to which the researcher would like to generalize her results, is those members of the general population who would take the new treatment

¹This bias goes by the name "randomization bias" in the literature on social experiments (Heckman (1992); Heckman & Smith (1995)).

²Another way to put this is that in randomized experiments, self-selection interferes with external validity but not internal validity. In non-randomized experiments or non-experiments, self-selection interferes with both external and internal validity.

if it were approved for sale.

My main findings are as follows. First, in the absence of a conventional alternative therapy that is available outside the trial, clinical trials do not suffer from self-selection. The reason is that all members of the population of interest would enroll in the trial.

Second, even if there is a conventional therapy, the only patients who engage in self-selection are those who believe the new treatment superior to conventional treatment and conventional treatment superior to no treatment. For these individuals, my primary finding is that, as the probability of randomization into active treatment (as opposed to placebo control) rises, the average effect of the new treatment on the trial population falls. The reason is that the higher the probability of treatment, the less optimistic a patient has to be about the new treatment (relative to conventional treatment) in order for her to prefer enrolling in a trial – a lottery between the new treatment and no treatment – over the non-trial alternative – certain consumption of conventional treatment. I verify this prediction with a data set of outcomes from over 120 trials of anti-ulcer medications with different probabilities of treatment.

An important implication of this finding is that self-selection causes the standard estimate of treatment effects – the difference between average outcomes in the new-treatment and placebo-control groups – to have a positive bias. The reason is that medical trials always have a probability of treatment less than one, whereas outside the trial context a patient who takes the

new treatment would do so with certainty, i.e., with probability equal one. If outcomes fall with higher probability of treatment, then outcomes in the real world must be smaller than outcomes in a trial. In my sample of ulcer trials, I estimate that selection bias accounts for roughly half of treatment effects for two medications (H₂-blockers and proton-pump inhibitors) and can reverse the sign of treatment effects for a third medication (prostaglandins).

My third finding is that an alternative experimental design which permits patients to choose among treatment lotteries can eliminate bias from self-selection. In this design patients are offered a choice between (a) a lottery over new treatment and placebo or (b) a lottery over new treatment and conventional control. All patients in the population of interest, i.e., who prefer new treatment, would enroll. Those that prefer conventional treatment to no treatment would simply enroll in the second lottery. The cost of this design is that one cannot identify the effect of new treatment relative to no treatment on these patients, but that information is not very valuable because these patients would take conventional therapy rather than no treatment if the new treatment were not available.

Although there is an extensive economics literature on self-selection across treatment and control groups in non-randomized studies, there is less discussion of self-selection across study participation and non-participation in the context of randomized experiments. That which exists notes that self-selection into a study is not a problem so long as one is interested only in the effect of treatment on the treated. If one is interested in treatment effects among a

broader population, however, the participation decision may introduce what is called randomization bias (Heckman (1992); Heckman & Smith (1995); Heckman (1996)). The limited literature on randomization bias does not model the behavior that generates this bias, as this paper does, but merely attempts to demonstrate empirically that such bias exists (Heckman (1992); Kamoinka & Lacroix (2002)).³ The literature on self-selection in non-randomized studies does not exhaust the analytic insights possible with randomized experiments because the lottery aspect of the latter permits greater theoretical structure to be placed on models of self-selection behavior. This reduces the work that must be done by data to identify the effects of selection.

In the "statistics for medicine" literature there is little discussion of patient self-selection into medical trials. Rather, the focus is on the selection of patients *by doctors* into trials, either via recruitment (see, e.g., Senore & Aimar (1999)) or exclusion criteria (see, e.g., Ellenberg (1994); Robinson & Lerner (1996)). Physician-selection might alter the impact patient self-selection has on estimates of treatment effects to the extent that it causes a deviation from random sampling from the population of interest. I leave that interesting question for future research. The few papers there are on the participation

³One ostensible exception is Philipson (2000), which examines the effect of changes in the proportion of a population that is treated on the treatment effect in that population. The distinction between that work and the present paper is twofold. First, in Philipson's paper, differences in the treatment probability change the treatment effect in the population (because of externalities of treatment on those not treated). In this paper, however, differences in the treatment probability change the *sampling of* treatment effects, not the treatment effects themselves. Second, Philipson is concerned with bias in estimates of the effect of treatment on the treated. I am concerned with the bias in estimates for a broader population.

decision by patients merely attempt to survey the reasons why patients choose not to participate in a medical trial (see, e.g., Verheggen & Jonkers (1998); Britton & Bain (1999)) or to document differences between the characteristics of patients who choose to enroll and those who do not (see, e.g., Olschewski & Davis (1992); Britton & Bain (1999)), without drawing conclusion about the impact of the differences on estimates of treatment effects.

The paper may be outlined as follows. Section 1 sets forth my model of self-selection into clinical trials. Section 2 test the predictions of my model against a data set of outcomes from ulcer trials. Section 3 proposes and evaluates an alternative trial design to estimate treatment effects without bias due to self-selection.

1 Model of Patient Self-Selection

1.1 Setup

Let us begin with some building blocks for the model. Assume an ill patient faces two possible future health states: continued illness \underline{y} or recovery \bar{y} .⁴ Utility in these states is $\underline{U} = u(\underline{y})$ and $\bar{U} = u(\bar{y})$, respectively. I will discuss three treatments for the patient's ailment, indexed by k : no treatment ($k = 0$), a new treatment ($k = 1$), and conventional treatment ($k = 2$). Let y_k be a random variable that gives the patient's health outcome following treatment k and define p_k as the probability of recovery with treatment k :

⁴The appendix extends the results of the paper to the case of continuous outcomes.

$p_k = \Pr \{y_k = \bar{y}\}$. Let π_k be a patient's belief about p_k . In general, π_k would be a distribution function, but it will make no difference to the analysis if it is a point estimate and so I shall assume that. In order for self-selection to affect estimates of treatment effects, it must be that there is heterogeneity of treatment effects and of beliefs about treatment effects in the population. Let $g(\mathbf{p}, \boldsymbol{\pi})$ be the joint probability distribution function for chances of recovery $\mathbf{p} = (p_0, p_1, p_2)$ and for beliefs $\boldsymbol{\pi} = (\pi_0, \pi_1, \pi_2)$ about these chances.

This last assumption requires some clarification and perhaps justification, at least with regard to the new treatment. I do not assume that patients know their response to new treatment, only that they have beliefs about the value of that treatment. These beliefs could be based on experience with related compounds, experience that the patients have previously responded well or poorly to treatments generally, news reports about the new treatment, advice from personal doctors, or information about the average value of all new treatments. The beliefs could be wrong – though I will rule out a particular type of wrong in subsection 1.3. While it would be useful to explore the source of beliefs about new treatment, it should not be controversial to assume that beliefs exist.

In order to formally model a medical trial, one needs to specify its design. My focus will be on blinded, randomized, placebo-controlled trials – the gold-standard according to the U.S. Food and Drug Administration. From the patient's perspective, the trial is a lottery over assignment to new treatment or the placebo. Let d be the probability of assignment to the new treatment arm.

This probability is revealed to patients during the informed consent process.⁵ The purpose of blinding is to prevent patients in the control arm from leaving or simultaneously seeking treatment outside the trial. In my analysis, blinded implies that researchers control the lottery payoffs that patients are offered. Without it, patients would, in effect, be able to pick the control arm payoff.

I assume that patients decide to enroll in a clinical trial if and only if their expected utility from the trial is greater than their expected utility from treatment outside the trial. To ensure that this choice has bite, I assume that the new treatment is only available through the trial. Otherwise no patients would enroll in trials without monetary compensation, which is generally limited to out-of-pocket expenses.⁶ I also assume that it is the utility of health and not the disutility of foregone consumption due to the cost of treatment that drives enrollment in clinical trials. This assumption may be justified by the aforementioned limit on monetary compensation and the fact that patients typically have health insurance that covers the marginal cost of care.

In order to fully characterize the patient's choice problem, note initially that expected utility from consumption of treatment k is

$$EU_k = \pi_k \bar{U} + (1 - \pi_k) \underline{U}$$

⁵Indeed, this is required by law in the U.S. See, 21 C.F.R. §§ 7.3(f), 20.25.

⁶Research institutions typically limit, by policy, compensation to levels not regarded as having an undue influence on the decision of patients to participate in research. See, e.g., University of North Carolina (2004), Pfizer (2005). Grady & Emanuel (2005) find that use of patient compensation is widespread but the amounts are modest. For more on the medical debate over compensation, see Dunn & Gordon (2005); Reiser (2005)

Expected utility outside the trial depends on whether the patient would choose no treatment or conventional treatment. Assuming that the patient chooses among these treatments on the same basis as she chooses whether to enroll in a trial, expected utility outside a trial is $\max\{EU_0, EU_2\}$.

Whereas outside the trial the patient has her choice of treatments, inside the trial she experiences a lottery over treatments. Because treatments themselves are lotteries over health states, the trial is actually a compound lottery over health states. Expected utility from the trial is

$$EU_T = \pi_T \bar{U} + (1 - \pi_T) \underline{U}$$

where $\pi_T = d\pi_1 + (1 - d)\pi_0$.⁷

The patient's choice between treatment within and without the trial will solve $\max\{EU_T, \max\{EU_0, EU_2\}\}$. This problem can be rewritten in an intuitive manner with a simple transformation of variables. Let $\tilde{p}_k = p_k - p_0$ be the health benefit of treatment k relative to no treatment, and $\tilde{\pi}_k = \pi_k - \pi_0$ be the patient's belief about this treatment k effect. Now the patient's problem becomes

$$\max\{d\tilde{\pi}_1, \max\{0, \tilde{\pi}_2\}\} \tag{1}$$

In other words, the patient's choice is between (a) a chance d at a treatment effect of $\tilde{\pi}_1$ with the new treatment, and (b) a certain treatment effect of 0

⁷This formulation is similar to that in Philipson & DeSimone (1997).

with no treatment or of $\tilde{\pi}_2$ with conventional treatment, whichever is greater.

An immediate implication of (1) is that risk aversion does not affect patient selection into medical trials. The reason is that outcomes are assumed to be binary. Therefore, if two lotteries offer the same expected outcome, they must also offer the same variation in outcomes. If health outcomes were continuous and one treatment offered a tighter distribution over outcomes than another, however, risk aversion might affect a patient's decision to enroll in a trial. That said, risk aversion would not, on the margin, always discourage enrollment in a trial. If the new treatment had lower outcome variance than conventional treatment or no treatment, a trial would attract the risk averse patient, *ceteris paribus*.⁸

1.2 Baseline

In order to determine whether and how self-selection introduces bias into estimates of treatment effects, one must specify the population of patients in which the researcher is interested. If the population is simply those that enroll in the trial, i.e., one is interested in the effect of "treatment on the treated," then it is well established that randomization eliminates self-selection bias. In my analysis, however, the population of interest is those patients who would take the new treatment if it were available outside the medical trial, i.e., patients who believe treatment effects of new treatment are superior to either no

⁸Harrison & Rutstrom (2005) fail to consider this possibility, and not surprisingly find that risk aversion does not lead to significant randomization bias in a Danish field experiment.

treatment or conventional treatment:

$$\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}^9 \tag{2}$$

The rationale is that it is these patients who are most likely to consume the new treatment if approved following the medical trial.¹⁰ Note that the trial population – those patients for whom $d\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}$ – is a subset of my population of interest. Everyone who enrolls in a medical trial meets the conditions for membership in the population of interest because $\tilde{\pi}_2/d > \tilde{\pi}_2$.

Some researchers may be interested in treatment effects among the whole population of patients. To these researchers, the ideal trial would enroll randomly selected patients from the population. In order to analyze bias from self-selection in this case, I recommend separating the population of interest into two subgroups: those who prefer new treatment (my population of interest) and those who do not. Total bias is the population-weighted average of bias in each of these groups. I will address selection bias in the group that does not prefer new treatment at the tail end of the next subsection.

⁹One objection to this definition is that the doctor and not the patient chooses the latter's treatment. But the same could be said of the study participation decision – the doctor determines whether the patient will enroll in a medical trial. In that case, one should simply substitute the treating physician's beliefs for the patient's belief in the analysis.

¹⁰I address the problem of learning in the conclusion.

1.3 Implications

Selection bias is defined to be the difference between treatment effects in the trial population and in the population of interest:

$$\begin{aligned} & E_g(y_1 - y_0 | d\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}) - E_g(y_1 - y_0 | \tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}) \\ &= [E_g(\tilde{p}_1 | d\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}) - E_g(\tilde{p}_1 | \tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\})] \Delta y \end{aligned}$$

where $\Delta y = \bar{y} - \underline{y}$. Bias arises from mismatches between the trial population and the population of interest (both of which are defined by patients' beliefs) that affect the sampling of treatment effects in a medical trial.

This description immediately highlights two cases where there is no bias from self-selection. The first is where there is no correlation between beliefs about treatment effects and actual treatment effects. Although there is self-selection based on beliefs, sampling of treatment effects remains random. The second case is where there is no conventional control. The condition for membership in the trial population and for membership in the population of interest converge to $\tilde{\pi}_1 > 0$. All patients who prefer the new treatment will enroll in a medical trial regardless of the probability of receiving it because the worst case outcome (placebo control) is as good as the only therapy available outside the trial (no treatment).¹¹

¹¹A related case immune to randomization bias is where there is a conventional treatment, it is the control treatment in the medical trial, and all patients prefer conventional treatment to no treatment. In this case, the worst outcome from the trial – conventional control – is at least as good as the perceived best treatment available outside the trial.

In reality, these conditions are uncommon, which is why self-selection warrants medical researchers' attention. In order to get theoretical traction on the bias from self-selection, I make two substantive assumptions. First, there is a positive correlation between a patient's beliefs about treatment effects and her actual treatment effects. More formally, (A1) $\tilde{p}_{ki} = f(\tilde{\pi}_{ki}) + \tilde{\varepsilon}_{ki}$, where i indexes individual patients, ε_{ki} can be thought of as prediction error by patients, $f' > 0$, $\varepsilon_{ki} \perp \pi_{ki}$, and $\varepsilon_{ki} \perp \varepsilon_{k'i'}$ for $k \neq k'$ or $i \neq i'$. The assumption of positive correlation is reasonable. It is consistent with patients under- or over-estimating the effects of treatment. The opposite case, a negative correlation between beliefs and outcomes, is implausible, implying that patients who respond well to treatment systematically believe treatment is worse than patients who respond poorly to treatment.

The second assumption (A2) is that the distribution of $\ln \tilde{\boldsymbol{\pi}}$, where $\tilde{\boldsymbol{\pi}} = (\tilde{\pi}_1, \tilde{\pi}_2)$, is log concave or log convex with mean $\boldsymbol{\mu}$ and variance $\boldsymbol{\Sigma}$. This is a non-restrictive assumption: the class of log-concave distributions is quite large and includes the multivariate normal. So, for example, if $g(\tilde{\boldsymbol{\pi}})$ were a right-truncated multivariate log-normal, the assumption is satisfied (Theorem 7, Bagnoli & Bergstrom (1989)).

Finally, I make the following "technical" assumption. Define $\tilde{\sigma}_k^2 = \text{var}(\tilde{p}_k)$, $\tilde{\sigma}_{12} = \text{cov}(\tilde{p}_1, \tilde{p}_2)$, $u_{ki} = \ln \tilde{\pi}_{ki} - \mu_i$, $W_i = u_{1i} - u_{2i}$, $\tilde{\sigma}^2 = \tilde{\sigma}_1^2 + \tilde{\sigma}_2^2 - 2\tilde{\sigma}_{12}$, $b_1 = (\tilde{\sigma}_1^2 - \tilde{\sigma}_{12}) / \tilde{\sigma}^2$, $b_2 = b_1 - 1$, and $V_i = b_1 u_{2i} - b_2 u_{1i}$. By construction $u_i = b_i W_i + V_i$, where W_i and V_i are uncorrelated. In my analysis, I go further and assume (A3) that W_i and V_i are independent.

These assumptions yield the central insight of this paper: so long as the correlation between the treatment effects of the new therapy and of the conventional therapy are not too great, an increase in the probability of assignment to new treatment will lower estimates of the treatment effect of the new therapy from the medical trial. More formally,

Proposition 1 *Under assumptions (A1)-(A3), if $\text{corr}(\tilde{p}_{1i}, \tilde{p}_{2i}) \leq \tilde{\sigma}_1/\tilde{\sigma}_2$, then $\partial E_g(\tilde{p}_1 | d\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\})/\partial d \leq 0$.*

The proof can be found in the appendix. The logic, however, is that the trial only attracts patients who believe that the benefit of the new treatment is so great that, even with a probability $1 - d$ of obtaining no treatment at all, the trial offers a better expected outcome than conventional treatment. As the probability of assignment to new treatment increases, the perceived benefit of the trial rises. Some patients who were not optimistic enough about the new treatment to have previously enrolled, will now do so. This will lower the average beliefs about the new treatment's effect among enrollees. Since expected treatment effects are monotonically increasing in beliefs, average treatment effects rise with average beliefs.

Critical to the proposition is the condition that the correlation between beliefs about the efficacy of the new and conventional treatments is not too high. If that is violated, then a higher probability of obtaining the new treatment may attract not just patients with more pessimistic beliefs about the new treatment, but also those with more optimistic beliefs about the new

treatment who nevertheless have not enrolled because they also have more optimistic beliefs about the conventional treatment. (Recall that patients enroll in the trial not merely because $\tilde{\pi}_{1i}$ is high, but because it is greater than $\tilde{\pi}_{2i}/d$.) Fortunately, the condition on the correlation of new and conventional treatment effects is rather lax. It is satisfied whenever the variance of beliefs about the new treatment is greater than the variance of beliefs about the conventional treatment, i.e., when $\tilde{\sigma}_1^2 \geq \tilde{\sigma}_2^2$, which is likely simply because the new treatment is new.

The economics behind Proposition 1 are identical to the Roy model (Roy (1951); Heckman & Honore (1990)) of sorting of workers across sectors. In that model, workers have sector-specific skills with associated skill prices. A worker chooses the sector that offers her the highest income, which is the product of skill and skill price. The effect of a change in one skill price on the distribution of workers across sectors, and thus the average skill level in each sector, depends on distribution of skills among workers and the correlation of skills in each sector. This maps neatly onto the problem of self-selection into clinical trials. Here there are two sectors: the trial and no trial. Skills are treatment effects. And the skill price in the trial is the probability of receiving the new treatment while the skill price outside the trial is one.

The practical import of Proposition 1 is that self-selection introduces positive bias into standard estimates of treatment effects. To see this, note that taking new treatment outside a trial is equivalent to a trial where the probability of receiving new treatment is one, and that the entire population of

interest would enroll in such a trial. Because treatment effects fall as the probability of treatment rises, it must be that consumption of the new treatment outside the trial (again, equivalent to a trial where the probability of treatment is one) must produce smaller treatment effects than new treatment in a trial (which generally has a probability of treatment below one). Indeed, a researcher could estimate the magnitude of bias if she had a number of trials with different probabilities of treatment. She would simply have to estimate the marginal effect of increases in the probability of treatment on estimates of treatment effects from individual trials and predict the treatment effect in a trial where the probability was one.

There are two circumstances in which selection bias may not be positive. First, if the medical trial offers significant monetary incentives for participation, then patients who prefer no treatment or conventional treatment may enroll.¹² In the former case, the bias is still positive because patients who prefer no treatment will have lower treatment effects than those who prefer new treatment, given the presumed positive correlation of beliefs and treatment effects. If patients who prefer conventional treatment take the bait, one cannot be sure treatment effects would be lower because there may be a number of patients believe that both new treatment and conventional treatment are really good, but that conventional treatment is just slightly better. These patients may raise the average treatment effect in the trial. In prac-

¹²Harrison & Rutstrom (2005) find that monetary compensation increases the average level of risk aversion among subjects in a field experiment.

tice, however, the number of such patients is likely to be small, especially if the correlation between new treatment and conventional treatment effects is not very large. Therefore, the selection bias will likely remain positive.

A second circumstance in which bias may not be positive is when the researcher cares about treatment effects in the general patient population. Now the baseline includes patients in my population of interest (those who prefer new treatment and for which selection bias is positive), plus all patients who prefer no treatment or conventional treatment. For the latter, the logic is similar to that for a monetary incentive. Among, in particular, those who prefer conventional treatment, the sign of bias is unclear because preferring conventional treatment may not imply that new treatment is less effective. So long as the correlation between new treatment and conventional treat is not too high, however, selection bias will remain positive.¹³

2 Application to Ulcer Trials

This section examines a data set of outcomes from 121 medical trials of three ulcer medications in order to test my model of patient self-selection. Proposition 1 provides the prediction to be tested: as the probability of treatment

¹³There is a literature that attempts to identify the full distribution of treatment effects, though in its parlance the goal is to identify the Roy model. The data requirements, however, are typically beyond the capabilities of medical researchers. For example, they would need knowledge about either the distribution of conventional treatment effects among those who do not enroll in the trial or about the proportion of the general patient population that enrolls in the medical trial (Heckman & Honore (1990)). They do not, however, even know locally the size of this population.

rises, estimated treatment effects should fall.¹⁴ The ulcer data confirm this prediction. Moreover, they suggest that selection bias is roughly half the size of treatment effects for two of the medications studied, and reverses the sign of treatment effects for the last medication.

2.1 Data

The first data set includes the published results from clinical trials studying treatment for non-gastric ulcers. Ulcers are the erosion of the mucous lining in the stomach or small intestine and are judged healed via endoscopy by the researcher. Three classes of medication are considered. The first class, H₂-blockers, was introduced in 1977, and is thought to prevent the production of acid in the stomach. The most popular brands are Tagamet (cimetidine), Zantac (ranitidine), and Pepcid (famotidine). The second class of medication, prostaglandins, was introduced in the mid-1980s and is thought to build up the mucous lining of the stomach and intestine. The most common prostaglandins are enprostil and misoprostil. The third class, proton-pump inhibitors (PPIs), were introduced after most prostaglandins, starting in 1989. Like H₂-blockers, these medications prevent the production of acid in the stomach. The most popular brands are Prilosec (omeprazole), Nexium (esomeprazole) and Prevacid (lansoprazole). A distinguishing feature of all three classes of medication

¹⁴If one knew whether the population distribution of treatment effects was log-concave or log-convex, one could also derive a prediction concerning the variance of treatment effects among the trial population (see Heckman & Honore (1990)). That distribution, however, is unknown.

is that they offer a much higher chance of healing an ulcer than do antacids or bismuth subcitrate, which are mainly palliatives.¹⁵

Each of the trials is randomized, double-blind, and parallel-armed, which means each patient is observed in only one treatment state. Trials employ either a placebo, antacid, bismuth subcitrate or conventional control. If conventional controls are employed, they are from a previous class of medication as the new treatment. PPIs are the highest class and H₂-blockers the lowest class of ulcer medication. Trials may have two or more groups (or arms). Typically there is one drug that is called the "new treatment" and a number that are called "controls," though if a control is simply a different dosage of the new treatment drug or a treatment of the same class of drugs as the new treatment drug, its arm is labelled in the data as a second new treatment arm for purposes of calculating the probability of receiving the new treatment.¹⁶

The probability of treatment is calculated by taking the number of arms with the most advanced class of new treatment and dividing by the total number of arms in the trial, on the theory researchers intend to assign the same number of patients to each arm to maximize the power of comparisons across arms.¹⁷ The fact that some trials examine multiple dosages of the

¹⁵That being said, it is now recognized that 90% of non-gastric ulcers are caused by the bacteria *heliobacter pylori*. These infections are usually treated with a combination of antibiotics and H₂-blockers or PPIs. This paper tests for placebo effects in trials where H₂-blockers, prostaglandins or PPIs are used in isolation. These trials typically predate the change to antibiotic-based treatments.

¹⁶While individuals may have different beliefs about efficacy of each experimental treatment versus the control, it is assumed that individuals do not have refined beliefs about relative efficacy of different treatments in the same class of medication, e.g., cimetidine v. ranitidine.

¹⁷Where the numbers actually randomized to two different arms of a trial differ from this

same new treatment or multiple treatments from same class of drug as the new treatment is what introduces variation in the probability of treatment. Table 1 summarizes, by the drug labelled new treatment in a trial, the number of *trials* with different types of control, numbers of treatment and control arms, and dates of publication. It also describes the frequency of *new treatment groups* by the type of control and the probability of receiving new treatment in the trial.¹⁸

Data were gathered on health outcomes and on the characteristics of trials and of patients. Individual patients' outcomes are binary: either a patient's ulcer was healed or it was not. The data do not, however, reveal individual patients' outcomes. Rather, they reveal patient outcomes aggregated to the treatment group-level. Thus, the observed outcomes are the share of patients whose ulcers had healed at a given date in each arm of each trial. Data on patients' characteristics are also aggregated to the treatment group-level. For example, there are data on the average age of patients assigned to any given treatment group, but not the age of each patient assigned to that group.

Trials take between one and four measurements each individual. These

by a factor of greater than 1.5, the probability of treatment is calculated by inferring the number of subjects intended to be randomized to each arm and dividing the number that were to be randomized to treatment arms and dividing by the total number that were to be enrolled in the trial.

¹⁸Note that no H₂-blockers trials in the sample employ conventional controls, prostaglandin trials are divided between placebo/antacid/bismuth subcitrate controls and conventional controls, and all PPI trials have conventional controls. This is mainly a product of the fact that H₂blocker trials typically preceded prostaglandin trials, which in turn typically preceded PPI trials. (There was one H₂-blocker trial with conventional control and one PPI trial without a conventional control. I removed these because, without a second trial, there is no possibility of variation in probability of treatment within these groups.)

measurements fall on one or more of the following dates: 1, 2, 3, 4, 6, 8, 10 or 12 weeks after onset of treatment. Different trials take different numbers of measurements and at different times. Table 1 also provides the frequency of trials by the number of measurements taken on the treatment arm, and the frequency of treatment arms, by the week of measurement. The table reveals that most trials take two measurements on each arm, typically at 2 and at 4 weeks. Although data on average patient characteristics for a group are available as of the date that patients are randomized into the group, precise information on how the group's characteristics changes over outcome measurement dates due to attrition are not available. With respect to outcomes, this omission is handled in three ways, by assuming those who attrite out heal at the same rate as those who remain (method 1), all heal (method 2), or all do not heal (method 3).

Table 2 provides summary statistics for the data. Each observation represents a measurement on the indicated arm of the indicated type of trial. Means and standard deviations are calculated weighting each arm in proportion to the number of patients evaluated per protocol at each measurement. Frequency of medication and total dosage are not provided for control arms because such variables are meaningless for placebo arms.¹⁹

¹⁹The antacid-permitted variable is coded from 1 to 5. One indicates that subjects were prohibited from taking antacids, two that subjects were discouraged from taking antacids, three that subjects were permitted to take antacids (or the study did not counsel subjects on antacids), four that antacids were provided, and five that antacids were required.

2.2 Empirical Strategy

I begin by assuming that the effects of treatment k on patient i enrolled in trial j is a linear function of a vector x_{ijk} , which includes a constant, observed clinical and demographic variables on the patient in treatment arm k , structural features of the trial, and an error term: $p_{ijk} = \beta'_k x_{ijk} + \varepsilon_{ijk}$. The error term captures variation due to unobserved characteristics of patients. I assume the errors are independent of x_{ijk} and are i.i.d. normal across individuals with mean zero and variance σ_k . This implies unobserved effects do not depend on the trial, but may depend on the treatment.

Because trials often take multiple measurements on each patient, I interpret a treatment's effect as a hazard rate. So p_{kijt} is the probability that an individual's ulcer will heal in period t given the effects of treatment k . Because the data only provide a snapshot of patient covariates at the beginning the trial, hazard rates are assumed to be constant and the t -subscripts on hazard rates and covariates are dropped. This implies $-\ln S_{ijk}(t)/t = \beta'_k x_{ijk} + \varepsilon_{ijk}$, where $S_{ijk}(t)$ gives the probability of still having an unhealed ulcer on date t . Summing over individuals and dividing by n_{jk} , the number of patients enrolled in treatment arm k of trial j , yields $-\overline{\ln S_{jk}(t)}/t = \beta'_k \bar{x}_{jk} + \bar{\varepsilon}_{jk}$. The left-hand side is first approximated by a first-order Taylor approximation around $\bar{S}_{jk}(t)$, the average probability of remaining ill at t . Then $\bar{S}_{jk}(t)$ is approximated by the observed group survival rate \bar{y}_{jkt} .²⁰ This last step is problematic because

²⁰Individual y_{ijkt} are binary variables that indicate whether the subject healed in period t . The observed group survival \bar{y}_{jkt} is the group-average of y_{ijkt} and is thus continuous in

data on patients who attrite out of the trials are not available. Therefore, \bar{y}_{jkt} is calculated under three different assumptions about the healing rate of those who attrite (methods 1 - 3 referenced earlier). The standard estimator of treatment response is the difference in average outcomes in the new treatment and control groups. Comparing groups k and k' at time t yields the equation $-\ln \bar{y}_{jkt} - \ln \bar{y}_{jk't} / t = \beta'_k \bar{x}_{jk} - \beta'_{k'} \bar{x}_{jk'} + \bar{\varepsilon}_{jk} - \bar{\varepsilon}_{jk'}$ where the unit of time is one day.

Self-selection is a problem only if it is not fully captured by \bar{x}_{jk} and the unobservable covariates that capture selection are correlated with \bar{x}_{jk} . Proposition 1 suggests that such selection is driven by the probability of treatment; specifically, it suggests that treatment effects should fall with this probability. To test this hypothesis I add d_j , the probability of treatment in trial j as a regressor. This yields the following regression equation:

$$-\ln \bar{y}_{jkt} - \ln \bar{y}_{jk't} / t = \beta'_k \bar{x}_{jk} - \beta'_{k'} \bar{x}_{jk'} + \gamma d_j + \bar{\varepsilon}_{jk} - \bar{\varepsilon}_{jk'} \quad (3)$$

Proposition 1 predicts that γ should be negative.²¹

Estimation is by feasible GLS procedures. Each observation is a measurement on a treatment group in a trial. Although trials typically take multiple measurements on each group, observations are weighted such that each group

[0, 1].

²¹Because the relevant hazard rate is confined by assumption to $[0, 1]$, the empirical model requires estimation of a linear probability model. While that model has flaws, it is not wholly inappropriate for the application in this paper. As a theoretical matter, the dependent variable in (3) can range from $(0, \infty)$. Moreover, because individuals in an arm are aggregated, the error term is more likely to resemble a normal distribution.

makes a contribution to estimates in proportion to the number of patients in the group, regardless of the number of measurements made on each patient. This weighting does not meaningfully alter the results. Standard errors are calculated assuming group-wise heteroskedasticity at the trial-level, but not at the arm-level. I only report estimates where the dependent variable is calculated assuming patients who attrite out heal at the same rate as those who are evaluated. Results from regressions which assume that those who attrite out either all heal or all do not heal do not materially differ. Four specifications of \bar{x}_j are employed. Specification (1) includes a constant, the probability of treatment d_j interacted with the indicators for the class of ulcer medication being tested and the nature of the control, an indicator for whether investigators excluded patients with "other serious problems" from the trial, and the interaction of this exclusion criteria and d_j ;²² (2) includes (1) plus trial level-variables (antacid role, daily frequency of medication, total daily dosage), with dosage interacted with indicators for the class of ulcer medication being tested and the nature of control employed; (3) includes (1) plus patient-oriented covariates (male, smoker, log age) for treatment groups (interacted with the class of new treatment) and control groups (interacted with whether the trial is conventional treatment-controlled); and (4) includes

²²Exclusion criteria are rules established by researchers to exclude certain types of patients from the trial. In industry-sponsored studies (at least 47% of the sample), the purpose is to eliminate treatment non-responders and elevate estimates of treatment effects. I include an indicator for exclusion criteria and an interaction with probability of treatment to account for this bias. (The results are robust to elimination of these variables.) The exclusion criteria I track is the rule against enrollment of patients with other serious problems, which is common in ulcer trials (63% of trials in the sample) as well in trials of other medications.

all of the above. The results are robust to standard modifications of these specifications, including exclusion of most interactions with d_j .

2.3 Results

Table 3 presents the core results from estimation of equation (3). The main finding is that coefficient estimates for interaction of trial indicators and the probability of treatment – reported in the middle section of panel B – are uniformly negative. Nearly all are statistically significant. This strongly supports Proposition 1.

In order to interpret the coefficient estimates, keep in mind, first, that the dependent variable is the difference in the average daily healing rate among patients in the treatment group and patients in the control group. So if the dependent variable is, say, .05, and there are 100 patients each in the treatment and the control groups, 5 more patients healed in the treatment group than in the control group. Second, the average medical trial assigns 50% of patients to new treatment. Therefore, my estimate of treatment effects in the typical H₂-blocker trials (under specification 1) is 0.017 (0.03, the coefficient on the H₂-blocker trial indicator, plus 0.5 times -0.026, the coefficient on the interaction between probability of treatment and the H₂-blocker trial indicator). Estimates of treatment effects in 50% trials for other medications and covariate-specifications are presented in Panel D. Third, predictions of treatment effects without selection bias are identical to predictions of treatment

effects in a trial where 100% of patients are assigned to new treatment. These estimates are reported in Panel E.

What we see is that PPIs tend to perform better than H₂-blockers and H₂-blockers sometimes better, sometimes worse than prostaglandins in 50% clinical trials without correction for selection bias. (Recall that PPI's are always being tested against conventional control.) When we compare treatments after correcting for selection bias, however, PPIs emerge as superior to H₂-blockers and H₂-blockers clearly better than prostaglandins. The reason is that selection bias was stronger in prostaglandin trials than in H₂-blocker or PPI trials. There are two explanations. First, because H₂-blockers were introduced before other classes of ulcer medications, there was no conventional alternative for patients outside of trials. Therefore, the entire population of interest probably sought to enroll. Second, all the PPI trials are conventional treatment (typically H₂-blocker) controlled. Therefore, there is a smaller downside with the trial. This remains true in the case of H₂-blocker controls because even selection-biased estimates of treatment effects suggested H₂-blockers were better than prostaglandins.

How large are my estimates of selection bias? The difference in biased and unbiased estimates of treatment effects is roughly half the size of treatment effects in 50% trials of H₂-blockers and PPIs. In trials of prostaglandins, however, the bias is roughly 2 to 6 times the size of treatment effects. In other words, in prostaglandin trials, self-selection appears to change the sign of treatment effects. A treatment that is worse than control actually appears

to perform better than control without a correction for selection bias!

3 Alternative Design

There is an alternative trial design that would permit the estimation of relevant treatment effects without bias from patient self-selection. Instead of offering patients a lottery over new treatment and *the investigator's choice* of control, placebo or conventional therapy, the design would offer a lottery over new treatment and *the patient's choice* of control, placebo or conventional therapy. (The alternative trial would offer the same probability of assignment to new treatment as the standard trial.) The two designs are illustrated in Figure 1.²³

The intuition behind the modified design is that the population of interest – defined to include only those who prefer the new treatment – includes two subgroups: those who prefer no treatment to conventional treatment and those who prefer conventional treatment to no treatment. The former will choose as their control placebo and the latter will choose conventional treatment.²⁴

Both subgroups, however, will enroll in the trial regardless of the probability

²³The alternative design is identical to offering both a placebo-control and a conventional-control trial to the same general patient population. The design is similar, but not identical, to an unblinded trial. There we assume the patient assigned to placebo control will seek conventional treatment outside the trial. But because the patient may not report this to the researcher, estimates of treatment effects relative to placebo will be polluted with estimates of treatment effects relative to conventional control.

²⁴In more formal terms, whereas the selection equation for a placebo-controlled trial is given by $d\tilde{\pi}_1 > \max\{0, \tilde{\pi}_2\}$, the selection equation for the alternative trial would be $\max\{d\tilde{\pi}_1, d\tilde{\pi}_1 + (1-d)\tilde{\pi}_2\} > \max\{0, \tilde{\pi}_2\}$, which collapses to definition of the population of interest (2).

of receiving new treatment, and estimates of treatment effects will not depend on this probability.

There are two problems with this design. First, the researcher will only be able to estimate the effect of new treatment relative to conventional treatment for the subgroup that prefers a conventional control. That information, however, is more valuable to this subgroup than an estimate of the effect of treatment relative to placebo. Second, the alternative design requires a larger sample size, and is therefore more expensive, than the standard placebo-controlled trial. The required sample size is not double, however, because the same outcomes in the new treatment arm can be used to calculate the effect of new treatment relative both to placebo and to conventional control. For example, in a parallel-armed trial, the sample size increment is only 50%.²⁵ The relevant question for the researcher, then, is whether selection bias is significant enough to warrant the extra cost of a larger sample.

²⁵Schouten (1999) derives the following approximation for sample sizes for a two-group trial where one group (n_2) is γ times as large as the other (n_1):

$$n_1 \geq (z_{1-\alpha/2} + z_{1-\beta})^2 \times \frac{(1 + \gamma) \sigma^2}{\gamma \delta^2} + \frac{z_{1-\alpha/2}^2}{2(1 + \gamma)}$$

where the groups are assumed to have the same variances σ^2 , α is the desired significance level, β is the desired power, and δ is the difference the researcher would like to detect. This equation implies that the total number of patients required with unequal groups is $(1 + \gamma)^2 / 4\gamma$ times the number required with equal group sizes. If the treatment group will be used for comparison to both controls, then $\gamma = 2$. Therefore, the total number of patients required is 9/8 that with equal group sizes. This means, instead of two trials with equal groups, the researcher will require one trial with 9/8 the size of an equal group trial and a second trial with just the smaller group, i.e., $9/8 \times 1/3$. Summing the two yields a factor of 3/2.

4 Conclusion

This paper not only suggests that medical trials are subject to what economists call randomization bias, but that this bias causes standard estimators to overstate treatment effects. The result is intuitive: experiments on a new treatment attract patients that are optimistic about the new treatment, and these patients probably respond better to that treatment. There is no reason to think the finding does not generalize to social experiments.

But the analysis also raises some more fundamental questions about evaluation of treatment effects. First, how do individuals form their beliefs about a new treatment, beliefs that are assumed to guide their decision whether to participate in a research study? Do they predict that new treatment has the same incremental benefit as the average medical innovation? Or do they simply experience Knightian uncertainty? If so, does decisionmaking under such uncertainty play out – with respect to study participation – like decisionmaking under risk? Second, how can one identify treatment effects in the relevant population in the presence of learning? The concern is that, if a study finds that new treatment has, say, a higher treatment effect than conventional treatment, people who previously preferred conventional treatment may now prefer the new treatment. The average treatment effect may be different in this larger population of interest than in the pre-study population of interest. This problem is an example of the principle that observation changes the system being observed.

A Appendix

This appendix provides the proof to Proposition 1 and extends it to the case of continuous outcomes.

A.1 Proof of Proposition 1

All expectations are taken with respect to g . An individual will choose to enroll if and only if $\tilde{\pi}_{1i} \geq \max\{0, \tilde{\pi}_{2i}/d\}$. Because

$$\begin{aligned} E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \max\{0, \tilde{\pi}_{2i}/d\}) &= \Pr\{\tilde{\pi}_{2i} \leq 0\} E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq 0) \\ &\quad + \Pr\{\tilde{\pi}_{2i} \geq 0\} E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d), \end{aligned}$$

$\text{sign}[\partial E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \max\{0, \tilde{\pi}_{2i}/d\}) / \partial d] = \text{sign}[\partial E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d) / \partial d]$. Because $\varepsilon_{ki} \perp \pi_{k'i'}$ for all (k', i') , $E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d) = E(f(\tilde{\pi}_{1i}) | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d) + E(\varepsilon_{ki})$. Because $f' > 0$, $\text{sign}[\partial E(\tilde{p}_{1i} | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d) / \partial d] = \text{sign}[\partial E(\tilde{\pi}_{1i} | \tilde{\pi}_{1i} \geq \tilde{\pi}_{2i}/d) / \partial d]$.

Since $g_{\tilde{\pi}}$ is log concave, $\tilde{\pi}_{ki}$ is a log-concave random variable, by Theorem 2 of Prekopa (1973). Given $\tilde{\pi}_{ki} > 0$, for $k = 1, 2$, $E[\tilde{\pi}_{1i} | \tilde{\pi}_{1i} > \tilde{\pi}_{2i}/d] = E[\tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d]$. Because $\ln x$ is monotone increasing in x ,

$$\begin{aligned} &\text{sign} \left\{ \frac{\partial E[\tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d]}{\partial d} \right\} \\ &= \text{sign} \left\{ \frac{\partial E[\ln \tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d]}{\partial d} \right\}. \end{aligned}$$

Because W_i and V_i are independent, $E[\ln \tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d] = \mu_1 +$

$b_1 E(W_i | W_i > c(d, \mu))$, where $c(d, \mu) = -(\mu_1 - \mu_2) - \ln d$.

Log-concavity or log-convexity of $(\ln \tilde{\pi}_{1i}, \ln \tilde{\pi}_{2i})$ implies log-concavity or log-convexity, respectively, of W_i , by theorem 5 in Bagnoli & Bergstrom (1989) and corollary 2 in An (1998), respectively. Propositions 1 and 2 in Heckman & Honore (1990) demonstrate that log-concavity or log-convexity of W_i implies $\partial E[W_i | W_i \geq c] / \partial c \geq 0$. Therefore,

$$\begin{aligned} & \frac{\partial E[\ln \tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d]}{\partial d} \\ = & \frac{\tilde{\sigma}_1^2 - \tilde{\sigma}_{12}}{\tilde{\sigma}^2} \frac{\partial E[W_i | W_i \geq c]}{\partial c} \frac{\partial c(d, \mu)}{\partial d}. \end{aligned}$$

Because $\partial c(d, \mu) / \partial d < 0$, $\partial E[\ln \tilde{\pi}_{1i} | \ln \tilde{\pi}_{1i} - \ln \tilde{\pi}_{2i} > -\ln d] / \partial d \leq 0$ so long as $\tilde{\sigma}_1^2 > \tilde{\sigma}_{12}$. This condition is the same as $\text{corr}(\tilde{p}_1, \tilde{p}_2) < \tilde{\sigma}_1 / \tilde{\sigma}_2$.

A.2 Continuous health outcomes

Suppose health is a continuous, univariate outcome, $y \in [\underline{y}, \bar{y}] = Y$, and that utility from health, $u(y)$, is strictly increasing in outcomes and constant across individuals. Define $p_{ki}(y)$ as the actual distribution of outcomes for patient i given treatment k and $\pi_{ki}(y)$ as patient i 's beliefs about the probability of each outcome. Let $g_{\tilde{p}}$ and $g_{\tilde{\pi}}$ give the population distribution of $\tilde{\mathbf{p}}_i = (\tilde{p}_{1i}, \tilde{p}_{2i})$ and $\tilde{\boldsymbol{\pi}}_i = (\tilde{\pi}_{1i}, \tilde{\pi}_{2i})$, where $\tilde{p}_{ki} = p_{ki} - p_{0i}$ and $\tilde{\pi}_{ki} = \pi_{ki} - \pi_{0i}$, for $k = 1, 2$.

The population of interest includes all individuals who satisfy

$$\int_Y u(y) \tilde{\pi}_{1i}(y) dy \geq \max \left\{ 0, \int_Y u(y) \tilde{\pi}_{2i}(y) dy \right\}$$

The trial population, however, includes only those for whom

$$\int_Y u(y) \tilde{\pi}_{1i}(y) dy \geq \max \left\{ 0, \frac{\int_Y u(y) \tilde{\pi}_{2i}(y) dy}{d} \right\}.$$

These conditions are analogous to the binary case.

Consider the following assumptions, the first three of which are slight modifications of (A1) - (A3).

A1') $g(\ln \pi_i | y)$ is nondegenerate either log-concave or log convex, with mean

$$\boldsymbol{\mu}(y) \text{ and variance } \boldsymbol{\Sigma}(y) \text{ for all } y \in Y.$$

Define $u_{ki} = \ln \tilde{\pi}_{ki} - \mu_i$, $W_i = u_{1i} - u_{2i}$, $\sigma = \sigma_{11} + \sigma_{22} - 2\sigma_{12}$, $b_1 = (\sigma_{11} - \sigma_{12}) / \sigma$, $b_2 = a_1 - 1$, and $V_i = b_1 u_{2i} - b_2 u_{1i}$. By construction $u_i = b_i W_i + V_i$, where W_i and V_i are uncorrelated.

A2') W_i and V_i are independent.

A3') $\int_Y y \tilde{p}_{ki}(y) dy = f(\int_Y y \tilde{\pi}_{ki}(y) dy) + \tilde{\varepsilon}_{ki}$, where where f is differentiable,

$$f' > 0, \tilde{\varepsilon}_{ki} \perp \tilde{\pi}_{k'i'} \text{ for all } (k', i'), \text{ and } \tilde{\varepsilon}_{ki} \perp \tilde{\varepsilon}_{k'i'} \text{ for all } (k', i') \text{ except } (k' = k, i' = i).$$

A4) $\int_Y y \tilde{\pi}_{ki} dy$ exists.

These yield the continuous analogue of Proposition 1.

Proposition 2 *Under assumptions A1' - A3' and A4, if $\tilde{\rho}_{12} \leq \tilde{\sigma}_1 / \tilde{\sigma}_2$, then*

$$\partial E(\tilde{p}_{1i} | s = BT) / \partial d \leq 0.$$

Proof. Suppressing the arguments of $u(y)$, $\tilde{p}_{ki}(y)$ and $\tilde{\pi}_{ki}(y)$ and the range of integration Y ,

$$\begin{aligned} E_g \left(\int y \tilde{p}_{1i} dy \mid \int u(y) \tilde{\pi}_{1i}(y) dy \geq \max \{0, (\int u(y) \tilde{\pi}_{2i}(y) dy) / d\} \right) \\ = \Pr \left\{ \int u \tilde{\pi}_{2i} dy \leq 0 \right\} E_g \left(\int y \tilde{p}_{1i} dy \mid \int u \tilde{\pi}_{1i} dy \geq 0 \right) \\ + \Pr \left\{ \int u \tilde{\pi}_{2i} dy > 0 \right\} E_g \left(\int y \tilde{p}_{1i} dy \mid \int u \tilde{\pi}_{1i} dy \geq \int u \tilde{\pi}_{2i} dy / d \right). \end{aligned}$$

Thus, $\text{sign}[\partial E_g(\int y \tilde{p}_{1i} dy \mid \int u \tilde{\pi}_{1i} dy \geq \int u \tilde{\pi}_{2i} dy / d) / \partial d] = \text{sign}[\partial E_g(\int y \tilde{p}_{1i} dy \mid \int u \tilde{\pi}_{1i} dy \geq \int u \tilde{\pi}_{2i} dy / d) / \partial d]$. Because u is strictly increasing, this is equal to $\text{sign}[\partial E_g(\int y \tilde{p}_{1i} dy \mid \int y \tilde{\pi}_{1i} dy \geq \int y \tilde{\pi}_{2i} dy / d) / \partial d]$. Because $f' > 0$ and $\varepsilon_{ki} \perp \pi_{k'i'}$, it is equal to $\text{sign}[\partial E_g(\int y \tilde{\pi}_{1i} dy \mid \int y \tilde{\pi}_{1i} dy \geq \int y \tilde{\pi}_{2i} dy / d) / \partial d]$.

By Prekopa (1973) Theorem 2, because $g_{\tilde{\pi}}$ is log concave, for any given $y \in Y$ and $k = 1$ or 2 , $\tilde{\pi}_{ki}(y)$ is distributed log-concave in the population. Pick any n members of Y and index them by j . Because the convolution of two log-concave random variables is also a log-concave random variable An (1996), $\sum_{j=1}^n (y_j/n) \tilde{\pi}_{ki}(y_j)$ is a log-concave random variable, for $k = 1, 2$. Because $\int_Y y \tilde{\pi}_{ki} dy$ is the limit of this sum as $n \rightarrow \infty$ and (by assumption) exists, $\int_Y y \tilde{\pi}_{ki} dy$ is log-concave, for $k = 1, 2$. By the logic of Proposition 1, then, $\text{sign}[\partial E_g(\int y \tilde{\pi}_{1i} dy \mid \int y \tilde{\pi}_{1i} dy \geq \int y \tilde{\pi}_{2i} dy / d) / \partial d]$ is negative. ■

Acknowledgments

The author thanks Gary Becker, Albert Choi, James Heckman, Feifang Hu, Steve Levitt, Willard Manning, Ron Michener, and Tomas Philipson, as well as workshop participants at the University of Virginia and the University of Chicago, for helpful comments. He also thanks David Greene, Maulshree Solanki, and Hema Srinivasan for their excellent research assistance.

References

- An, Mark Yuying. 1996 (June). *Log-concave Probability Distributions: Theory and Statistical Testing*. Centre for Labour Market and Social Research, University of Aarhus, Working Paper 96-01.
- An, Mark Yuying. 1998. Logconcavity versus Logconvexity: A Complete Characterization. *Journal of Economic Theory*, **80**, 350–369.
- Bagnoli, Mark, & Bergstrom, Ted. 1989. *Log-Concave Probability and Its Applications*. University of Michigan.
- Britton, Annie., Martin McKee Nick Black Klim McPherson Colin Sanderson, & Bain, Chris. 1999. Threats to applicability of randomised trials: exclusions and selective participation. *Journal of Health Services Research and Policy*, **4**(2), 112–121.
- Dunn, Laura B., & Gordon, Nora E. 2005. Improving Informed Consent and

- Enhancing Recruitment for Research by Understanding Economic Behavior. *Journal of the American Medical Association*, **293**(5), 609–612.
- Ellenberg, J. H. 1994. Selection Bias in Observational and Experimental Studies. *Statistics in Medicine*, **13**(5-7), 557–567.
- Grady, Christine, Neal Dickert Tom Jawetz Gary Gensler, & Emanuel, Ezekiel. 2005. An analysis of U.S. practices of paying research subjects. *Contemporary Clinical Trials*, **26**, 365–375.
- Harrison, Glenn W., Morten Igel Lau, & Rutstrom, E. Elisabet. 2005. *Risk Attitudes, Randomization to Treatment, and Self-Selection into Experiments*. Centre for Economic and Business Research Discussion Paper.
- Heckman, James, & Honore, Bo E. 1990. The Empirical Content of the Roy Model. *Econometrica*, **58**(5), 1121–1149.
- Heckman, James J. 1992. *Randomization and Social Policy Evaluation*. Harvard University Press. Pages 201–230.
- Heckman, James J. 1996. Randomization as an Instrumental Variable. *Review of Economics and Statistics*, **78**(2), 336–341.
- Heckman, James J., & Smith, Jeffrey A. 1995. Assessing the Case for Social Experiments. *Journal of Economic Perspectives*, **9**(2), 85–110.
- Kamoinka, Thierry, & Lacroix, Guy. 2002. *Assessing the Extent of Randomization Bias in the Canadian Self-Sufficiency Experiment*. Working paper.

- Olschewski, Manfred, Martin Schumacher, & Davis, Kathryn B. 1992. Analysis of Randomized and Nonrandomized Patients in Clinical Trials Using the Comprehensive Cohort Follow-up Study Design. *Controlled Clinical Trials*, **13**, 226–239.
- Pfizer. 2005. *Policy on Compensation to Human Research Subjects*. Available at http://www.pfizer.com/pfizer/are/about_public/mn_about_ethical_trials_human.jsp.
- Philipson, Tomas. 2000. *External Treatment Effects and Program Implementation Bias*. NBER Working Paper No. 0250.
- Philipson, Tomas, & DeSimone, Jeffrey. 1997. Experiments and Subject Sampling. *Biometrika*, **84**, 221–234.
- Prekopa, Andras. 1973. On Logarithmic Concave Measures and Functions. *Acta Scientiarum Mathematicarum*, **33**, 335–343.
- Reiser, Stanley Joel. 2005. Research Compensation and the Monetization of Medicine. *Journal of the American Medical Association*, **293**(5), 613–614.
- Robinson, Delbert, Margaret G. Woerner Simcha Pollack, & Lerner, Gail. 1996. Subject Selection Biases in Clinical Trials: Data from a Multicenter Schizophrenia Treatment Study. *Journal of Clinical Psychopharmacology*, **16**(2), 170–176.
- Roy, A. D. 1951. Some Thoughts on the Distribution of Earnings. *Oxford Economic Papers (New Series)*, **3**, 135–146.

- Schouten, Hubert J. A. 1999. Sample Size Formula with a Continuous Outcome for Unequal Group Sizes and Unequal Variances. *Statistics in Medicine*, **18**, 87–91.
- Senore, Carlo, Renaldo N. Battista Antonio Ponti Nereo Segnan Stanley H. Shapiro Stefano Rosso, & Aimar, Daniela. 1999. Comparing Participants and Nonparticipants in a Smoking Cessation Trial: Selection Factors Associated with General Practitioner Recruitment Activity. *Journal of Clinical Epidemiology*, **52**(1), 83–89.
- University of North Carolina, Office of Clinical Trials. 2004. Developing Good Recruitment Practices. *Office of Clinical Trials Newsletter*, **2**(4), 1, 3.
- Verheggen, Frank W.S.M., Fred Nieman, & Jonkers, Ruud. 1998. Determinants of patient participation in clinical studies requiring informed consent: Why patients enter a clinical trial. *Patient Education and Counseling*, **35**, 111–125.

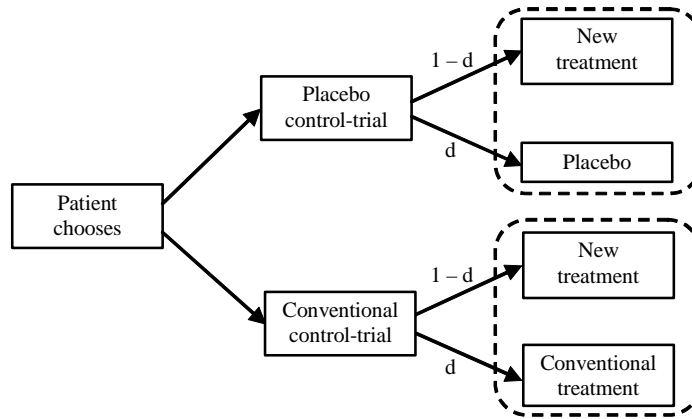


Figure 1: Alternative "choose your control" design.

Table 1: Frequencies of trials and treatment groups by design characteristic.

Frequency of trials	Treatment			Total	Frequency of groups	Treatment			Total
	H2-blocker	Prosta-glandin	PPI			H2-blocker	Prosta-glandin	PPI	
Total	69	24	28	121	Total	119	52	86	257
Type of control					Type of control				
Placebo, antacid or bis. subcitrate	69	14	0	83	Placebo, antacid or bis. subcitrate	119	29	0	148
Conv. treatment	0	10	28	38	Conv. treatment	0	23	86	109
Number of treatment groups					Probability of receiving treatment				
1	61	20	21	102	0.25	3	0	0	3
2	6	4	6	16	0.33	0	0	2	2
3	2	0	1	3	0.43	0	0	4	4
Number of control groups					0.50	90	40	44	174
1	66	24	27	117	0.67	12	12	30	54
2	0	0	1	1	0.75	14	0	6	20
3	3	0	0	3	Date of measurement (weeks)				
Number of measurements					1	4	0	2	6
1	39	6	0	45	2	21	16	36	73
2	25	14	20	59	3	2	0	2	4
3	5	4	7	16	4	63	25	36	124
4	0	0	1	1	6	19	8	3	30
Date trial results were published					8	7	1	7	15
1975	18	0	0	18	10	1	0	0	1
1980	30	5	0	35	12	2	2	0	4
1985	16	18	11	45					
1990	4	1	13	18					
1995	1	0	2	3					
2000	0	0	2	2					

Table 2: Summary statistics.

Variable	Treatment								
	H2-blockers			Prostaglandins			PPI		
	Grps	Mean	SD	Grps	Mean	SD	Grps	Mean	SD
Share healed:									
Treatment group	119	0.659	0.210	52	0.640	0.198	86	0.791	0.182
Control group	118	0.438	0.250	52	0.589	0.300	86	0.654	0.221
Probability of treatment	119	0.570	0.117	52	0.546	0.075	86	0.536	0.099
Exclusion criteria (no other serious problems) (0/1)?	119	0.730	0.446	52	0.610	0.492	86	0.713	0.455
Antacid permitted in trial (1-5)?	119	3.46	0.83	52	3.64	0.57	86	3.24	1.09
Frequency of dosage (times/day)	118	2.48	1.03	52	2.64	0.94	86	1.18	0.66
Total daily dosage (mg/1000)	118	0.5078	0.4407	52	0.0854	0.2258	86	0.0272	0.0180
Male (0/1)?									
Treatment group	110	0.760	0.089	47	0.728	0.095	84	0.701	0.082
Control group	110	0.762	0.089	47	0.738	0.107	84	0.686	0.081
Ever smoke (0/1)?									
Treatment group	84	0.611	0.111	49	0.572	0.135	80	0.482	0.116
Control group	84	0.624	0.097	49	0.558	0.121	80	0.483	0.150
Age (log years)									
Treatment group	102	3.83	0.11	47	3.81	0.15	76	3.81	0.11
Control group	102	3.83	0.10	47	3.80	0.15	76	3.81	0.12

Notes. Each observation is a measurement on a treatment group in a trial. Observations are weighted such that each group makes a contribution to estimates in proportion to the number of patients in the group, regardless of the number of measurements made on each patient.

Table 3: Estimates of selection bias and corrected treatment effects.

	(1)	(2)	(3)	(4)
A. Covariate specification				
Trial characteristics (dosage, frequency, antacid use)		x		x
Individual characteristics (sex, smoking, age)			x	x
B. Sample size				
Number of arms	145	144	110	110
Number of arms x measurements (= total observations)	252	251	200	200
C. Coefficients				
H2-blocker v. placebo trial indicator (p-value)	0.030 (0.0000)	0.028 (0.0060)	0.193 (0.0010)	0.182 (0.0030)
Prostaglandin v. placebo trial indicator	0.063 (0.0000)	0.058 (0.0000)	0.125 (0.1600)	0.035 (0.7400)
Prostaglandin v. conventional control trial indicator	0.144 (0.0000)	0.127 (0.0000)	0.279 (0.0100)	0.374 (0.0000)
PPI v. conventional control trial indicator	0.052 (0.0000)	0.035 (0.0070)	0.102 (0.3310)	0.231 (0.0000)
Probability x H2-blocker/placebo trial	-0.026 (0.0460)	-0.026 (0.1160)	-0.063 (0.0000)	-0.037 (0.0120)
Probability x Prostaglandin/placebo trial	-0.085 (0.0000)	-0.075 (0.0000)	-0.144 (0.0000)	-0.131 (0.0030)
Probability x Prostaglandin/conv. treatment trial	-0.317 (0.0000)	-0.296 (0.0000)	-0.439 (0.0000)	-0.484 (0.0000)
Probability x PPI/conv. treatment trial	-0.037 (0.1040)	-0.021 (0.3570)	-0.085 (0.0010)	-0.015 (0.4800)
Probability x Exclusion criteria	0.052 (0.0000)	0.050 (0.0030)	0.093 (0.0000)	0.067 (0.0000)
Exclusion criteria (other serious problem) indicator	-0.023 (0.0020)	-0.020 (0.0310)	-0.049 (0.0000)	-0.034 (0.0020)
D. Treatment effect in 50% trial				
H2-blocker v. placebo	0.017 (0.0000)	0.014 (0.0000)	0.162 (0.0030)	0.163 (0.0070)
Prostaglandin v. placebo	0.020 (0.0000)	0.021 (0.0000)	0.053 (0.5350)	-0.030 (0.7730)
Prostaglandin v. conventional treatment	-0.014 (0.0000)	-0.021 (0.0000)	0.060 (0.4590)	0.132 (0.0360)
PPI v. conventional treatment	0.034 (0.0000)	0.024 (0.0000)	0.060 (0.5560)	0.224 (0.0000)
E. Corrected treatment effect (predicted in 100% trial)				
H2-blocker v. placebo	0.004 (0.5220)	0.001 (0.8790)	0.130 (0.0170)	0.145 (0.0160)
Prostaglandin v. placebo	-0.022 (0.0030)	-0.016 (0.1270)	-0.019 (0.8190)	-0.095 (0.3620)
Prostaglandin v. conventional treatment	-0.172 (0.0000)	-0.170 (0.0000)	-0.160 (0.0180)	-0.110 (0.0730)
PPI v. conventional treatment	0.015 (0.1640)	0.013 (0.2530)	0.017 (0.8610)	0.216 (0.0010)
F. Difference (due to selection bias)				
H2-blocker v. placebo	-0.013 (0.0460)	-0.013 (0.1160)	-0.031 (0.0000)	-0.018 (0.0120)
Prostaglandin v. placebo	-0.042 (0.0000)	-0.037 (0.0000)	-0.072 (0.0000)	-0.065 (0.0030)
Prostaglandin v. conventional treatment	-0.158 (0.0000)	-0.148 (0.0000)	-0.219 (0.0000)	-0.242 (0.0000)
PPI v. conventional treatment	-0.018 (0.1040)	-0.011 (0.3570)	-0.042 (0.0010)	-0.007 (0.4800)

Notes. This table reports results from estimation of equation (3) by feasible GLS. Each observation is a measurement on a treatment group in a trial. Observations are weighted such that each group makes a contribution to estimates in proportion to the number of patients in the group, regardless of the number of measurements made on each patient. Standard errors are calculated assuming group-wise heteroskedasticity at the trial-level, but not at the arm-level. The dependent variable is $-\ln \bar{y}_{jkt} - \ln \bar{y}_{jk't}]/t$, where \bar{y}_{jkt} is the fraction of patients who heal in trial j of treatment k at time t , measured in days. The dependent variable is calculated assuming patients who attrite out heal at the same rate as those who are evaluated. Four specifications of \bar{x}_j are employed. Specification (1) includes a constant, the probability of treatment d_j interacted with the indicators for the class of ulcer medication being tested and the nature of the control, an indicator for whether investigators excluded patients with "other serious problems" from the trial, and the interaction of this exclusion criteria and d_j ; (2) includes (1) plus trial level-variables (antacid role, daily frequency of medication, total daily dosage), with dosage interacted with indicators for the class of ulcer medication being tested and the nature of control employed; (3) includes (1) plus patient-oriented covariates (male, smoker, log age) for treatment groups (interacted with the class of new treatment) and control groups (interacted with whether the trial is conventional treatment-controlled); and (4) includes all of the above.