CLINICAL TRIALS WITH NON-ADHERENCE & UNBLINDING: A GRAPHICAL PERSPECTIVE

ELIZABETH SILVER

Abstract. Clinical trials are often complicated by both nonadherence and unblinding. This makes it difficult to accomplish the main purpose of the trial: discover the causal effect of taking the treatment. I develop static and time-series graphical representations of the causal structure of trials, in order to represent how nonadherence interacts with unblinding to cause bias. I briefly describe some ways trials could be designed in order to avoid, or at least better evaluate the risk of such bias. I then compare intent-to-treat, per protocol, and instrumental variables analyses in terms of the causal assumptions they make regarding non-adherence, unblinding and dropout. I argue that all of them are unbiased when the blind is successful, but all of them may be biased when the blind fails (though not all to the same degree). I describe how Judea Pearl's 'front door' method is unbiased in the presence of unblinding, although it makes other strong unverifiable assumptions. I then consider the possible effects of measurement error on the front door method, per protocol, and instrumental variables. I show that using the front door method, the causal effect of treatment is identifiable in linear models, if we make two measurements of each error-prone variable. Unfortunately per protocol analyses are incompatible with linear models, implying that either other identifying parametric assumptions must be found, or that adherence should be measured as accurately as possible, using electronic monitors. Lastly, I argue that the scientific and regulatory communities' preferences for intention-to-treat analyses over per protocol is based on misinterpretation of historical results from the Coronary Drug Project, and a degree of skepticism that is not applied consistently to all sources of bias.

Contents

Part 1. Introduction

The Evidence-Based Medicine (EBM) movement holds randomised controlled double-blind trials to be the gold standard for evaluating new medical treatments. [5, 15] When EBM proponents justify this standard, they gloss over the problems that arise in real trials, and cite inferential properties that can only be said to hold of *ideal* trials. [2] An ideal trial would look something like this:

The Ideal Trial: A sample of participants is randomly assigned to two groups, A and B. One of the two groups is given the new treatment, and the other is given the control (either a placebo or an alternative, better-understood treatment). Neither the patients, nor the doctors treating them, know who is getting which treatment. The participants all take the treatment assigned to them, and at the end of the trial their medical outcomes are recorded and compared (again by physicians unaware of who was taking what). Any difference between the two groups' outcomes is then attributed to the difference between the treatments.

In the abstract, the logic of the Ideal Trial is unassailable, and if that were all there were to it, medical research would be considerably easier. The following examples illustrate how actual trials deviate from the ideal, and each one illustrates a different facet of the problem – and some surprising opportunities.

1. Motivating examples

1.1. **Example 1: The WHI study of Calcium and Vitamin D for Osteoporosis.**

In 2004, the New England Journal of Medicine published the results of the Women's Health Initiative (WHI) trial of Calcium & Vitamin D for Osteoporosis. [25] The aim of the trial was to show that a cheap, over-the-counter supplement could reduce the number of fractures in postmenopausal women; as the treatment offered no large profits to pharmaceutical companies, the trial was funded by the NIH. It included 36,262 women, and follow-up ran for an average of 7 years – a monumental achievement.

The results of the trial, however, were less impressive. Here is the Conclusion section from the abstract, the "bottom line":

> Among healthy postmenopausal women, calcium with vitamin D supplementation resulted in a small but significant improvement in hip bone density, did not significantly reduce hip fracture, and increased the risk of kidney stones.

That sounds definitive. If a trial of thirty-six thousand women could not detect a statistically significant effect, it is probably not clinically significant either.

However, the discussion section of the paper offered several plausible alternative explanations for the null result. Most women in the trial were already getting adequate dietary calcium, and only a few were deficient in Vitamin D at baseline, so the supplement could not have a strong beneficial effect in these women (although it might give them kidney stones). Furthermore, the dose of vitamin D (400 IU) had been shown, in other studies, to have only a weak effect; only doses of 600 IU or more had shown strong effects on fracture risk. Also, fractures were rare: fewer than half the predicted number of fractures actually occurred, drastically reducing the study's statistical power.

Lastly: over seven years of follow-up, many patients stopped taking the pills. By the end of the trial, 24% of participants admitted that they had completely stopped adhering, and only 59% of participants were still taking 80% or more of their medication (as measured by returned pill count, which tends to overestimate actual adherence). As US Surgeon General C.E. Koop put it, "Drugs don't work in patients who don't take them". In the WHI the converse appeared to be true: calcium and vitamin D did work in women who took them. In the Reply to Comments, the authors themselves wrote:

> Nonetheless, we believe that the trial results provide several indications that calcium intake does reduce the risk of hip fracture. Calcium and vitamin D supplementation reduced the risk of hip fracture by 29 percent among women with an adherence of 80 percent or more, 21 percent among those 60 years of age or older at enrollment, and 30 percent among those not taking other calcium supplements during the trial (all 95 percent confidence intervals for the correspond- ing hazard ratios exclude 1). In fact, we believe that these data support current recommendations for adequate calcium intake.

It's curious that the subgroup results were not included in the abstract, whereas the intent-to-treat[1] null result served as the main analysis, despite good reasons to think that it underestimated the effect of calcium and vitamin D. Subgroup results are often thought to be susceptible to bias, but in this case the main analysis itself was suspect. The question of this thesis is, "how should we analyse randomised controlled trials?" That also means, "if we do more than one analysis, how should we emphasise the different results, and which should we take to be primary?" Which analyses are taken to be more important on the grounds that they are more reliable?

1.2. **Example 2: Blind v. Unblinded Assessments of Multiple Sclerosis.** Noseworthy et al. [35] reported the results of a Canadian cooperative trial of cyclophosphamide, prednisone, and plasma exchange for multiple sclerosis. But this was not just a trial of the treatment; it was also a trial of the methodology.

Most randomised controlled trials are "double-blind", meaning that neither the patients, nor the doctors assessing their outcomes, know who is in which group. The Canadian trial's outcome measure was score on the Expanded Disability Status Scale (EDSS), which must be assessed by a neurologist. Noseworthy et al. tested the effect of blinding the assessors. At each checkup, every patient was assessed by two neurologists: one blind to the patient's assignment, and one unblinded.

In fact, the unblinded neurologists gave the patients in the treatment arm better scores on the EDSS than the patients in the placebo arm. The blinded physicians gave placebo and treatment patients identical scores – if anything, they slightly favoured the placebo group.[2] This example demonstrates that the results of trials can be biased if outcome

[1] See Section 11.1 for a description of intent-to-treat analyses.

[2] The study also measured whether participants and physicians could guess the treatment allocation. In fact, participants could guess their allocation significantly better than chance, but the blinded physicians were unable to do so. Interestingly, despite being able to guess better than chance, participant guess was not associated with outcome. This indicates that unblinded participants are not guaranteed to unblind their assessors, although of course it is not a proof that they can *never* unblind their assessors.

assessors are unblinded. The 'beneficial effect' reported in the unblinded arm had nothing to do with the treatment administered, only the assessors' *expectations* of a benefit.

1.3. **Example 3: Dose Timing for HAART.** Despite causing problems by reducing study power, non-adherence in the trial can also offer opportunities that are rarely exploited. The third example is not from a RCT at all, but it illustrates the potential advantages to pursuing these methods in a trial context.

Liu et al. [32] report the results of a longitudinal study of Highly Active Anti-Retroviral Therapy (HAART) for HIV. In particular, the researchers attempted to characterise the effect of *dose-timing errors* (DTEs) on viral load, and on viral resistance to HAART. The rationale: HAART has a short half-life, so variation in dose timing will cause serum concentration of HAART to drop repeatedly into a dangerous zone where it is insufficient to suppress viral replication, but sufficient to exert a strong selective effect on the virus. [52] Thus, more so than with other medications, we would expect DTEs to be associated with outcomes.

This is exactly what Liu et al. found. Even after controlling for percentage of doses taken, dose timing errors were significantly predictive of both viral load and viral resistance. Note that these DTEs would never be picked up with traditional "pill count" measures of adherence; they are only detectable using electronic monitoring devices, which record when the patient opens the drug bottle.

HAART is not the only treatment for which particular patterns of adherence have specific effects. Any drug to which people can become tolerant, any drug with first-dose effects, and any drug with rebound effects will have a particular adherence-related profile; and the profile will change with different formulations of the drug that have different half lives or absorption rates. Thus, large clinical trials with nonadherence represent an opportunity to learn these adherence-related effect profiles early in drug development, and provide advice to patients on what aspects of adherence are most important, what to do if they miss a dose, etc. However, this opportunity can only be exploited if adherence is measured in sufficient detail, which means using electronic monitoring devices. [50, 51]

2. Questions trials should answer

It should now be clear, given the three examples, that there are several possible research questions we might expect clinical trials to answer. The FDA is particularly interested in one question, for approving new drugs:

(1) Superiority (or equivalence): Is the treatment under study superior to a placebo (or at least as good as the current standard of care)?

For simplicity, I will limit my discussion to placebo-controlled trials.

Besides the superiority question, policy-makers, doctors and patients are often interested in *how much* better the treatment is. We also need to specify *who* it is better for. For

There are numerous other examples of participants becoming unblinded, [10, 40] but I know of no other experimental tests of how blinding affects the conclusion of a trial.

example, if we were considering withdrawing a treatment that was already in widespread use, we would want to know:

(2) Effect of Treatment on the Treated (ETT): Among the sub-population of people who are already taking the treatment, how different would their outcomes have been if they had all taken the placebo instead of the treatment?

The people currently taking the treatment may have had unusually good experiences with it, so the effect it has on them might be very different from the average effect it has on the whole population.

However, if we are considering introducing a new drug, we do not know in advance who will start taking the treatment. In that case we want to know the effect of the drug on the entire clinical population:

(3) Average Causal Effect (ACE): What would the average difference in outcomes be if we gave everyone in the clinical population the treatment, as opposed to giving them all the placebo?

Because new drug development is common whereas removing drugs from the market is rare, I will focus on the ACE instead of the ETT throughout the rest of this thesis.

So far I have not distinguished clearly between *being prescribed* treatment and actually *taking* it. Policy-makers are interested in the effect of prescribing the treatment to people, and so are doctors, but patients want to know what will happen if they take the medicine. So there is another pair of distinct questions:

(4) (Average Causal) Allocation Effect: What would the average difference in outcomes be if we allocated everyone in the clinical population to the treatment group, as opposed to allocating them all to the placebo group?

The effect of being prescribed treatment (averaged over people who adhere perfectly, and people who make mistakes or fail to adhere at all) is sometimes called "use-effectiveness" or "population efficacy".

However: When estimating the allocation effect, we want to rule out the reporting bias demonstrated by Noseworthy et al. [35]. Noseworthy found that allocating patients to the active treatment improved their outcomes if they were assessed by an unblinded neurologist, but not if they were assessed by a blinded neurologist (see Section 1.2). Allocation had an effect, but not because the treatment was effective. Thus, even when we are estimating the effect of allocation on outcome, we don't want the *total* effect – we just want to learn what portion of the effect of allocation on outcome is mediated by the treatment (see also Section 2.1). So (4) should really be:

(4b) Allocation Effect Mediated by Treatment: Of the average difference in outcomes caused by allocating people to treatment, versus allocating them to placebo, what portion of that difference is mediated by them receiving treatment?

This is a composition of the causal effect of allocation on treatment, and the causal effect of treatment on outcomes.[3]

Patients, however, want to know the effect of simply taking treatment – that is to say, the direct effect of treatment on outcome:

(5) Treatment effect: What is the average difference in outcomes we would see if we made people take the treatment, versus making them take placebo?

The effect of taking the treatment exactly as prescribed, compared to taking the placebo exactly as prescribed, is sometimes called "method-effectiveness".

Patients who frequently make mistakes also want to know the effects of those mistakes. There are several different ways patients can adhere imperfectly: missing doses, taking "drug holidays" (i.e. 3+ days of missed doses), making dose timing errors, discontinuing treatment too early, or overdosing, to name a few.

For a given type of non-adherence – say, a missed dose – there are then two contrast cases. My outcome given I missed a dose could be compared to the outcome I'd have had if I'd taken that dose. This kind of contrast tells us about potentially dangerous consequences of particular patterns of non-adherence. It puts an upper bound on what we can achieve if we intervene on adherence – for example, with a public health education campaign – among patients who are taking the treatment.

(6a) Adherence effect (a): What is the average difference in outcomes we would see if we made people miss one dose of treatment, versus making them take that dose?

Alternatively, my outcome after missing a dose could be compared to the outcome I'd have had if I had missed doses of the placebo. This would tell us whether the treatment is helpful even given imperfect adherence.

(6b) Adherence effect (b): What is the average difference in outcomes we would see if we made people miss one dose of treatment, versus missing one dose of placebo?

Clearly, contrast (a) and contrast (b) can be applied to the various different forms of non-adherence.

On the face of it, it looks like non-adherence within the trial would make it hard to estimate the Treatment Effect, but would actually help to estimate the various Adherence Effects, and the Allocation Effect. This is because the clinical population will include some non-adherent patients, just like the trial population.

However, trialists often want an affirmative answer to the superiority question – "is this treatment better than a placebo?" – in order to get regulatory approval. (The exceptions are trials like the WHI, which are not funded by pharmaceutical companies.) As a result,

[3] We can express this counterfactual contrast precisely using the do-calculus:

$$\sum_{out}\sum_{t} out \times P(Outcome = out|do(Treatment = t))P(Treatment = t|do(Allocation = 1))$$

$$-\sum_{out}\sum_{t} out \times P(Outcome = out|do(Treatment = t))P(Treatment = t|do(Allocation = 0))$$

trialists sometimes exclude patients thought to be poor adherers, for example by implementing pre-trial "run-in" procedures. [38] Excluding some non-adherers from the trial, or strongly encouraging adherence, means that the effect of treatment (as measured by an intention-to-treat analysis) will be less diluted by non-adherence. As a result it will be closer to the true Treatment Effect – but further from the true Allocation Effect, which is relevant to policymakers.

To learn the Allocation Effect we need an additional assumption: that rates of adherence in the trial represent rates of adherence in the clinical population. This assumption is often violated because trialists want a positive superiority result. This trade-off illustrates some of the tension between the different goals of trials – tension between the "explanatory" and "pragmatic" approaches to clinical trials. [42]

When I discuss different methods for analyzing the results of trials, I will point out which question(s) each method is supposed to answer. Ideally, for any given trial, we could perform several different analyses and answer all our questions. However, the design of each trial will still favor some questions over others.

2.1. "Open" (unblinded) trials and the placebo effect.

One might say I've ignored an important question: What is the effect of taking a treatment, and knowing you're taking it? Call this the Open Label Effect. It is what we intend to measure when we run an unblinded or "open label" trial – a trial in which all participants know what treatment they have been assigned to.[4] The Open Label Effect combines the physiological effects of the treatment with the "placebo effect", the psychological effects of knowing that you're taking treatment. There is evidence that psychological effects can have important physiological effects, [26][5] so the Open Label Effect would seem to be worth studying, as it most closely resembles what will happen in clinical practice, where no-one is blinded.

On the other hand, several arguments are traditionally made in favour of blinding, and these all count against investigating the open label effect. The claim is that the effect measured in open label trials may be biased. It may be influenced by factors other than true effects of the treatment.

There are two general categories of biases that can arise from unblinding: *reporting* biases and *differential treatment* biases. These effects arise from unblinding of up to six different groups of people: participants, healthcare providers, data collectors, outcome assessors, data analysts, and the data safety and monitoring committee. [8]

2.1.1. *Reporting biases.*

These arise from saliency/expectation effects and from attribution effects. Patients and outcome assessors are more likely to notice the effects they expect to see, and to overlook the unexpected. This would be especially true of outcomes that are not always reported (e.g. "other" category adverse effects) or those that are subjective (difficult differential diagnoses, pain, etc) compared to non-subjective outcomes that are

[4] Sometimes this is the only kind of trial we can run, because the treatment cannot be concealed. Major surgery, exercise, diet regimens, and use of eye-patches for amblyopia ('lazy eye') are all examples of treatments that are impossible to blind.

[5] Although the placebo effect alone may be impossible to estimate independently of various confounding factors. [28]

always reported (e.g. death). [11, 41] Patients and outcome assessors are also more likely to attribute what they notice to known possible causes. For example, even if two patients notice the same variation in their condition – say a headache – the patient who knows she is taking an active treatment is more likely to attribute the headache to the treatment than the patient who knows she is taking the placebo. There is no point reporting the headache to trial staff if she knows it cannot be an adverse effect. [28]

2.1.2. *Differential treatment biases.* These occur when knowledge of assignment influences what happens to the patient. Patients may decide to stop adhering to a placebo who would have adhered to active treatment (or vice versa), or they may drop out differentially, or use different co-interventions. Healthcare providers may recommend different co-treatments, differentially adjust the dosage, or encourage different behaviour (adherence, dropout, etc) in each group. They may differentially withdraw patients from the trial, or simply communicate different attitudes towards each of the two groups. [41] Data analysts may decide which analysis to do, and which participants to exclude from the analysis, based on their assignment. Trial staff may spend more energy pursuing participants lost to follow-up from one of the two groups. [11]

Thus, there are several reasons to think that an effect we are interested in – that of taking the treatment while knowing that you are taking it – may be contaminated by reporting biases, and by differential treatment of the two groups. The degree to which the trial's results differ from the effect seen in clinical practice depends on the strength of these effects.

This strength is difficult to estimate, and is likely to depend on the particular treatment and outcome in question. A meta-analysis by Schultz et al. (1995) found that trials described as double-blind reported effects 17% smaller than trials not so described. However, the trial reports did not indicate whether the blind was held successfully, and some of the trials may have been blinded despite not being described as such. [40] Furthermore, that meta-analysis took pairs of trials from a variety of different sub-fields of medicine (all within the Cochrane Pregnancy and Childbirth database), and the particular fields included may have influenced the result. Nevertheless, Schultz et al.'s result is reason to think these biases are worth worrying about. Although we may be interested in estimating the effect of knowing that you are taking the treatment, we are more motivated to prevent these biases from influencing our estimates of efficacy. [6]

[6] There are also reasons why we might not bother to learn the open-label effect at all, if we have already learned the pure treatment effect. Firstly, the placebo response component is likely to be unstable. Two hundred years ago, if a patient was given leeches to cure a fever, they probably experienced a positive psychological effect, because leeches were thought to be a useful remedy. Today, any patient given leeches for a fever should experience alarm and distrust. As medical knowledge changes, the placebo response will change, whereas the physiological effects of the treatment should remain relatively stable. In fact, given that the treatment's effects are unknown during the trial, the placebo effect within the trial may be quite different to the placebo effect seen in clinical practice. (The physiological responses will also change, as the general health of the population changes. For example, two hundred years ago in north America, a greater proportion of people had intestinal parasites than currently have them in the USA today, and fewer

3. Outline of the thesis

The basic question I'm asking is, "Which analyses do the best job of answering the questions we're interested in, given how real trials differ from the ideal version?" The answer depends on what we think is going on in clinical trials.

In Part 2 I will develop a graphical representation of the causal processes operating in a clinical trial with non-adherence, and potential unblinding. This representation should allow us to compare different analytic approaches in terms of which causal pathways influence their results. I'll also take a historical detour, explaining how the Coronary Drug Project has been misunderstood.

In Part 3 I'll describe different ways to deal with the bias caused by unblinding. These will range from actual solutions (measuring more variables, estimating the strength of bias, etc) to simply assuming the problem doesn't exist (e.g. when performing an intent-to-treat, per protocol, or instrumental variables analysis).

In Part 4 I'll consider what happens to our best methods when we measure variables imprecisely. Linearity and multiple measurements will save 'front door' and instrumental variables analyses, but not per protocol.

In the Conclusion I will argue that the research community's preference for intention-to-treat analyses is based on a degree of skepticism that is not applied uniformly to all sources of bias, and make recommendations for the design and analysis of clinical trials.

Part 2. Experimental Design and Analytical Assumptions: Nonadherence, Unblinding, and the Causal Structure of Real Trials

All analyses make certain assumptions in order to get a result. Often these assumptions are grounded in the design of the trial – for example, assigning patients to groups at random, by design, supports the assumption that no demographic factor influences a patient's assignment. However, some assumptions are more convenient than well-founded. When the assumptions are violated, the results of the analysis can be mistaken; ineffective or unsafe treatments may be given regulatory approval (or safe, effective treatments passed over); and patients pay with their lives.

As a result, arguments for preferring one analysis over another frequently hinge on which analysis' assumptions are better justified, and which are more likely to lead to serious mistakes if they are violated. For example, Russell Katz explains the FDA's preference for intention-to-treat analyses like so:

> "[...] the Food and Drug Administration does have great interest in seeing the results of an intent-to-treat analysis. The Food and Drug Administration view is that such an analysis has the great advantage of not introducing

suffered from allergies. However, much of our physiology remains the same, whereas medical knowledge has undergone a massive transformation.)

Secondly, the placebo effect may be achievable by other means – for example, by taking a sugar pill that is labeled as the active treatment – which are considerably cheaper and less risky than active treatment. Thus, even if the placebo effect is beneficial, it is not necessarily a unique advantage of the treatment. Alternative treatments (or even non-treatment) might be given the same advantage.

> the kind of bias that postrandomization exclusions can produce. Such biases can lead to erroneous conclusions, as shall be seen. Intent-to-treat analyses are not without problems, but they do provide critical information. Clearly, the Food and Drug Administration feels that the intent-to-treat analysis is the standard against which other analyses must be compared; after all, it is the only analysis that the Food and Drug Administration specifically requires to be included in all new drug applications. But, again, it bears repeating that it is not necessarily the *only* analysis that may be judged to be acceptable." [27, page 252]

I describe intent-to-treat analyses in Section 11.1. Here, I aim to illuminate the kinds of biases Katz is referring to. To that end, I will discuss at length the causal processes that operate in clinical trials, and how those processes justify or prohibit the kinds of analytic assumptions we might like to make. This section is supposed to give a causal interpretation to the assumptions and aid your intuitions regarding them.

4. Assumed background knowledge

4.1. **Causal graphical models.** My approach is to translate the design features and assumptions into the causal graphical models framework whenever possible,[7] so they can be compared easily. Providing a general introduction to causal graphical models is beyond the scope of this thesis, and the reader is referred to several excellent articles and books on the topic, from most accessible to most comprehensive: [13], [36], [43, part III], [37], [46]. I will assume general familiarity with:

- the concept of a "direct cause" relative to a set of variables
- the "direct effect" and "total effect" of some variable A, relative to a graph and to at least two possible values for A
- the concepts "path", "directed path" and "undirected path", and "back-door path"
- a variable being a collider (or non-collider) relative to a given path
- d-separation properties of graphs, and the resulting conditional independence constraints
- occasionally I will mention the Causal Markov Condition and the do-calculus.

All diagrams of graphical models are assumed to be *common-cause complete* – that is, of the set of variables included in the graph, no two of them share a common cause that has been omitted from the graph. Following convention, I draw circles around unobserved variables.

5. The causal structure of a clinical trial

Say we want to learn the on-treatment effect (i.e. the effect of actually taking treatment, as opposed to merely being assigned to take it). We need a general representation of the clinical trial, so we can compare the different statistical techniques for learning this effect.

[7] Parametric assumptions, for example, are not represented in this framework and require separate treatment.

Figure 1 represents the basic causal structure of a randomised, double-blind clinical trial. All the features of this causal diagram are drawn from our background knowledge:

(1) *Allocation* is exogenous[8] because patients are allocated to groups by randomisation.
(2) We know that *Allocation* influences whether a patient receives *Treatment* or not.
(3) We know (or at least can't rule out) that *Treatment* influences *Outcome*.
(4) Because the trial is blinded, the only way that *Allocation* can influence *Outcome* is through *Treatment*.
(5) We allow the possibility that some unidentified, unobserved causes **U**[9] influence both adherence to *Treatment*, and *Outcome*.

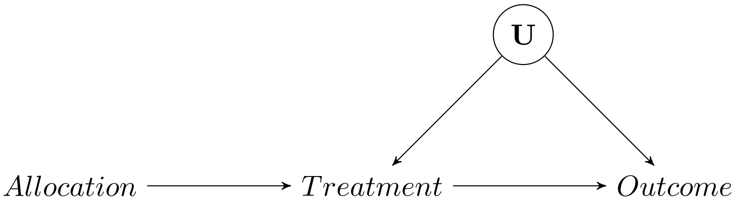

FIGURE 1. The causal structure of a randomised, double-blind trial with non-adherence

By contrast with Figure 1, Figure 2 represents an *unblinded* trial. In Figure 2, there is a direct effect of *Allocation* on *Outcome*. This might occur if reporting bias or differential treatment (see Section 2.1) affected patients' measured clinical *Outcome* without going through *Treatment*.

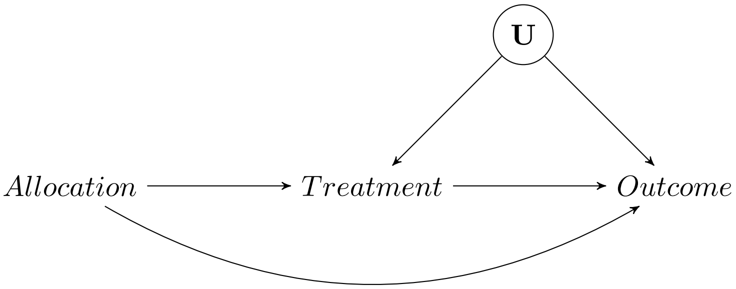

FIGURE 2. The causal structure of an unblinded trial

[8] I.e. it has no parents.

[9] **U** is circled in Figure 1 to follow the convention of circling unobserved variables. The bold typeface reminds us that **U** may represent a set of several variables. The structural relationships among the members of **U** would matter if we were trying to (a) block the path through **U** by conditioning on them, or (b) learn more about the causal structure by observing the *d*-separation relationships between members of **U** and other variables. So long as we are not trying to learn about **U** and are assuming that the path through **U** is always open, we can represent **U** simply as a single node in the graph.

Our problem is now clear. We want to find the effect of *Treatment* on *Outcome*, but there is a back-door pathway through **U**, which cannot be blocked by conditioning because **U** is unobserved. (If the trial is unblinded, then we have a second back-door pathway through *Allocation*, but *Allocation* is observed, so we can block that path.)

If we wished to find the effect of *Allocation* on *Outcome* that goes through *Treatment*, we would have no problems if the trial were blinded; if it were unblinded, then the direct effect of *Allocation* on *Outcome* would bias our analysis. (See Section 2.1 for why we would prefer to exclude the *Allocation* → *Outcome* edge.)

However, this representation is too basic to represent much of the literature; it does not even represent *Adherence* as a variable. To represent more analyses, and to find a solution to our problem, we need to expand our representation.

6. Non-Adherence

In the real world, patients do not always receive the treatment they were assigned to take, for many reasons (including forgetfulness, resolution of the illness, serious side effects, etc.). This is why *Allocation* and *Treatment* are separate variables in Figures 1 and 2. But many analyses refer to *Adherence* instead of *Treatment*. These are related but distinct variables. *Adherence* refers to the patient's dosing behavior;[10] as a result, patients assigned to take a placebo can have 100% adherence, if they take the placebo religiously. *Treatment*, a.k.a. *Treatment Received*, refers to the amount of active treatment received; so patients assigned to placebo automatically receive zero treatment (unless they somehow obtain it in violation of the study protocol).

In the literature, authors typically use only one of these two variables, and the choice depends on their purpose. Causal analyses use *Treatment* as the variable of interest, because adherence to placebo is assumed to have no causal effect. Associational analyses, however, usually use *Adherence* as the variable of interest, because even if it has no causal effect, adherence behavior is associated with other favorable prognostic factors. People who take their placebo religiously tend to do other healthy things, too. Per protocol analyses, for example, compare placebo-adherers with treatment-adherers, excluding non-adherers from both groups.

To understand and compare the causal and associational analyses, we need to represent both *Adherence* and *Treatment* in our framework. Figure 3 expands our graph to represent both variables.

It is crucial to realise that ***Treatment* is a deterministic function of its two parents, *Adherence* and *Allocation*.** If I know which group a patient was allocated to, and I know how many of the prescribed pills she took, then I know exactly how many pills of the active treatment she took. That is why there is no edge from **U** to *Treatment*; nothing besides *Allocation* and *Adherence* can influence *Treatment*. Consequently, if I condition on *Allocation* and *Adherence*, I effectively condition on *Treatment* as well,

[10] There are also cases where 'adherence' is partly determined by the physician – for example, a physician might decide a patient is too sick for surgery, or requires a higher dose of the medication, etc. Whether adherence is defined as being under patient control or not depends on the study.

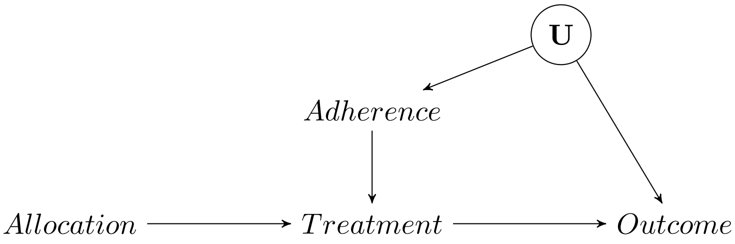

Figure 3. Expanded causal structure representing both Adherence and Treatment in a blinded trial.

so the conditional independence relationships are somewhat different from those of regular *d*-separation. For a formal characterisation of some of the properties of graphs with deterministic relationships, see [46, pages 53–56].

Again, the features of the graph are drawn from our background knowledge. Note that:

(1) There is no edge from *Allocation* to *Adherence*; this model assumes that group assignment has no influence on adherence. This depends on blinding. It might be false if, for example, one of the treatments has unpleasant side-effects. (See Section 7 for relaxation of this assumption.)
(2) There is no edge from *Adherence* to *Outcome*. This represents our assumption that simply taking pills does not influence medical outcome, except via the active treatment contained within them.
(3) As in the previous example, there is an unobserved (set of) variable(s), **U**, that is a common cause of both *Adherence* and *Outcome*. This could stand for variables like conscientiousness, stress, health-literacy, etc. that we would expect to influence both adherence and medical outcome.

Given this structure, we can see that per protocol analyses (see Section 11.2) are unbiased: condition on *Adherence*, and you block the back–door pathway between $Treatment$ and $Outcome$. It's a very simple measure, but assuming the graph in Figure 3 is correct, it should work.

So why does the FDA prefer intention-to-treat analyses (see Section 11.1) over per protocol? Because Figure 3 may not represent the true graph. If the trial is unblinded, *Allocation* may influence *Adherence*, so conditioning on *Adherence* opens up a pathway through **U** (see Figure 4). The FDA wants to prevent this pathway from biasing the results of the trial.

On the other hand, if the trial is unblinded, *Allocation* may have a direct effect on $Outcome$, so the intention-to-treat analysis may also be biased! The strong preference for intention-to-treat analyses seems strange, given that in the same circumstances when we'd expect per protocol analyses to be biased – namely unblinded trials – we would also expect that intention-to-treat analyses might be biased.

6.1. The Coronary Drug Project & the argument against using adherence data.

I believe that historical events helped make the FDA, and the biostatistics community

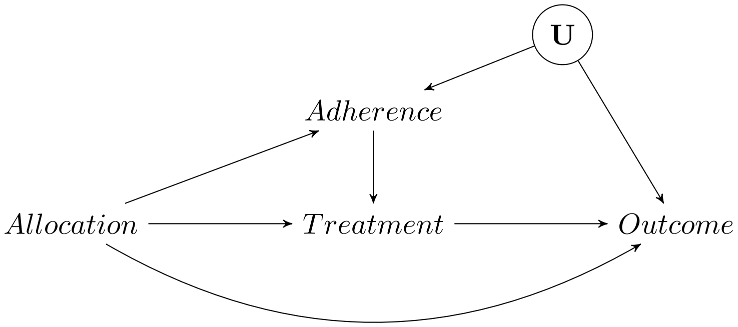

Figure 4. Causal structure of a Trial with Non-Adherence and Unblinding

in general, sensitive to bias through the *Allocation* → *Adherence* ← **U** → *Outcome* pathway. One particular trial, the Coronary Drug Project (CDP), contributed to this extreme wariness. In the CDP, the *Adherence* ← **U** → *Outcome* path produced a very strong association between *Adherence* and *Outcome* (see Table 1). The trial authors used this effect to argue that bias could result from any analysis that conditioned on *Adherence* – even though in the CDP, it appeared that the *Allocation* → *Adherence* edge was either absent or extremely weak; and even though conditioning on *Adherence* made no difference to the trial results.

Table 1. Death rates within the CDP adherence subgroups

	Clofibrate	Placebo
Adherance $\geq$ 80%	15.0%	15.1%
Adherance $<$ 80%	28.2%	24.6%
Total	20.0%	20.9%

The history of arguments against per protocol analyses is clearly a tangent from our investigation into the causal structure of a clinical trial. However, there are two reasons to cover it. Firstly, CDP was very influential. The Cochrane Collaboration's[11] Handbook for Authors of Systematic Reviews [22] cites it as a cautionary tale, and Russell Katz refers to it when he explains the FDA's reservations regarding per protocol analyses. [27] Secondly, I believe the CDP is not a cautionary tale at all, and the fact that it has been interpreted as one reflects a confusion in the literature. It's important to understand this confusion, so that the arguments in this thesis do not seem trivial or obvious. Apparently they are not obvious to the FDA or the Cochrane Collaboration.

[11] The Cochrane Collaboration is an international non-profit dedicated to performing and maintaining high-quality meta-analyses of medical research. Without meta-analyses, the volume of medical research would be too massive for individual doctors to follow. The Cochrane Collaboration, having no vested interests, is extremely influential.

Here's how the Cochrane Handbook describes the CDP:

> **'As-treated' (per-protocol) analyses**
>
> [...] A similarly inappropriate approach to analysis of a study is to focus only on participants who complied with the protocol. A striking example is provided by a trial of the lipid lowering drug, clofibrate (Coronary Drug Project Research Group 1980). The five-year mortality in 1103 men assigned to clofibrate was 20.0%, and in 2789 men assigned to placebo was 20.9% (P=0.55). Those who adhered well to the protocol in the clofibrate group had lower five-year mortality (15.0%) than those who did not (24.6%). However, a similar difference between 'good adherers' and 'poor adherers' was observed in the placebo group (15.1% vs 28.3%). Thus, adherence was a marker of prognosis rather than modifying the effect of clofibrate. These findings show the serious difficulty of evaluating intervention efficacy in subgroups determined by patient responses to the interventions. Because non-receipt of intervention can be more informative than non-availability of outcome data, there is a high risk of bias in analyses restricted to compliers, even with low rates of incomplete data. [22, `http://handbook.cochrane.org/chapter_8/8_13_2_2_high_risk_of_bias_due_to_incomplete_outcome_data.htm`]

It's is not at all clear how the results of the CDP demonstrate "a high risk of bias in analyses restricted to compliers". Comparing clofibrate-compliers with placebo-compliers produces exactly the same result as comparing the whole clofibrate group with the whole placebo group; clofibrate has no effect. If anything, the CDP seems almost miraculous in how little association there is between *Allocation* and *Adherence* (an independence you'd only expect to see in a perfectly blinded trial).

Russell Katz from the FDA provides an excellent summary of the CDP Research Group's original argument:

> The authors of the study attempted to determine if the differences in the two placebo subgroups could be established through the use of multivariate statistical methods. A multiple linear regression analysis of 5-year mortality and compliance was carried out on 40 baseline characteristics used as adjusting variables. This analysis yielded an adjusted mortality for good compliers of 16.4% and 25.8% for poor compliers. The minimal differences between the unadjusted and adjusted mortality rates demonstrate that differences in these baseline variables account for little of the differences in outcome between the subgroups defined by compliance. Clearly, then, other fundamental differences between the patients in the subgroups have not been detected by this detailed analytic procedure. What this implies is that compliers are, in some unknown and perhaps unknowable way, different from noncompliers.
>
> The authors also attempted to determine if patients who complied with the prescribed clofibrate regimen were benefited, despite the absence of any

overall effect on mortality. They concluded that there was no clearly valid way to arrive at this conclusion, primarily because it was impossible to identify the appropriate placebo comparison group. For example, the 5-year mortality for poor clofibrate compliers was 24.6% but was only 19.4% for all placebo patients. Alternatively, mortality in the good clofibrate compliers was only 15.0%, as compared with 19.4% for the placebo patients.[12] However, the argument can be made that the use of all placebo patients as the control group is inappropriate, because it is known that the two compliance subgroups have dissimilar outcomes.

One then could compare the between-treatment responses within compliance strata (i.e., clofibrate good compliers compared with placebo good compliers, and the corresponding poor compliers), which would give still different results. The authors suggest that any conclusion could be justified, depending on which groups are compared, but that any of the results of these subgroup comparisons are unreliable, because there is absolutely no assurance that the compliers in the placebo group are the same as the compliers (or noncompliers to noncompliers) on both known and unknown factors that might affect outcome. It is possible, for example, that the reasons for compliance (or noncompliance) are different between treatment groups and that those differences might have an effect on the outcome. It has been seen that at least in this case, detailed statistical manipulation was unable to detect any systematic reason for the differences between the two subgroups of compliers, yet clearly differences must exist. The only way to be assured that groups are comparable in all relevant respects (i.e, on those factors that may affect outcome), both known and unknown, is through the randomization process. Excluding patients after randomization, however "logical" the maneuver appears to be, can result in groups unbalanced on (unknown) critical factors, thereby introducing bias into the trial. Analyzing all patients randomized to treatment allows the assumptions on which

[12] Unbelievable as it seems, Katz is being charitable. The CDP research group really did make this argument (references to tables omitted):

> "Other analyses indicate additional difficulties in interpreting data on adherence and mortality. The five-year mortality was 24.6 per cent for poor adherers in the clofibrate group, as compared with 19.4 per cent for all patients, regardless of adherence, in the placebo group. On the other hand, mortality in good adherers in the clofibrate group was substantially lower than mortality in the placebo group (15.0 vs. 19.4 per cent) [...]. However, it may be argued that combining the two adherence subgroups of the placebo groups in such an analysis is almost certainly inappropriate, since the two subgroups have such dissimilar mortality results. If the adherence subgroups for clofibrate are compared with the corresponding subgroups for placebo, the five year mortality in poor adherers to clofibrate is lower than that in poor adherers to placebo (24.6 vs. 28.2 percent), whereas there is no difference in mortality between good adherers in the clofibrate group and good adherers in the placebo group (15.0 vs. 15.1 per cent) [...]. Therefore, one can justify almost any conclusion, depending on the analysis chosen." [20]

> the successful interpretation of the data rely to be fulfilled and hence gives meaning to significance testing. [27]

That's the argument. The CDP research group published zero empirical evidence that "the reasons for compliance (or noncompliance) are different between treatment groups", but there was no in-principle guarantee that *Allocation* had no effect on *Adherence*.

Despite the lack of *a priori* guarantee, there was empirical evidence that *Allocation* had no effect on *Adherence*. Firstly, the distribution of adherence was similar between the clofibrate and placebo arms (see Figure 5). A Pearson chi-square test for a difference between the clofibrate and placebo adherence distributions is not even remotely significant: $\chi^2(5) = 5.86$, $p = 0.32$. Likewise, a Pearson chi-squared test for the independence of *Allocation* and *Outcome* conditional on *Adherence* is non-significant: $\chi^2(3) = 1.89, p = 0.60$. The distribution of side effects shows only a very slight predominance of side effects for clofibrate (see Figure 6, and ignore the Niacin columns).

So how did the Coronary Drug Project become the textbook example for cases when *Allocation* could influence *Adherence*, such that conditioning on *Adherence* would introduce bias through the *Allocation* → *Adherence* ← **U** → *Outcome* pathway? The authors (rightly) emphasized the strength of the association between *Adherence* and *Outcomes*. Given the *Adherence* ← **U** → *Outcome* pathway was so strong, they realized that even a weak *Adherence* → *Outcomes* edge could cause substantial bias. However, they only attempted to "control for" causes along the *Adherence* ← **U** → *Outcome* pathway, and made no investigation of whether there was an *Allocation* → *Adherence* edge in their trial. Instead, they argued that in principle, this edge was always possible.

The CDP's paper presented the epitome of an unbiased per protocol analysis, and then claimed that it might be biased. This set a standard for unbiasedness that no actual trial could surpass. As a result, the CDP suppressed the use of adherence data for exploratory analyses. Because they did not investigate the presence or mechanism of an *Allocation* → *Adherence* edge, they did nothing to encourage researchers to think clearly about the causal processes producing bias, and ways in which it could be identified, measured and controlled.

In 1980 the CDP research group wrote that "there is no way of ascertaining precisely how or why the patients in the clofibrate and placebo groups have selected themselves or have become selected into the subgroups of good and poor adherers." [20] The CDP is still cited today and its arguments are repeated without qualification. [7, page 374–5] [22] [27, page 671] [33, page 254–5] We can do better, I argue, if we think clearly about the mechanism that might make participants adhere for different reasons in one group than in the other: namely, unblinding.

7. Blinding

Blinding is a design feature of most randomised controlled trials. A trial is called 'double-blind' if neither the patients, nor the physicians evaluating their outcomes, are told the patients' group assignment. However, I want to distinguish between the *procedure* of blinding and its intended *effect*, which is to keep knowledge of the assignment from influencing

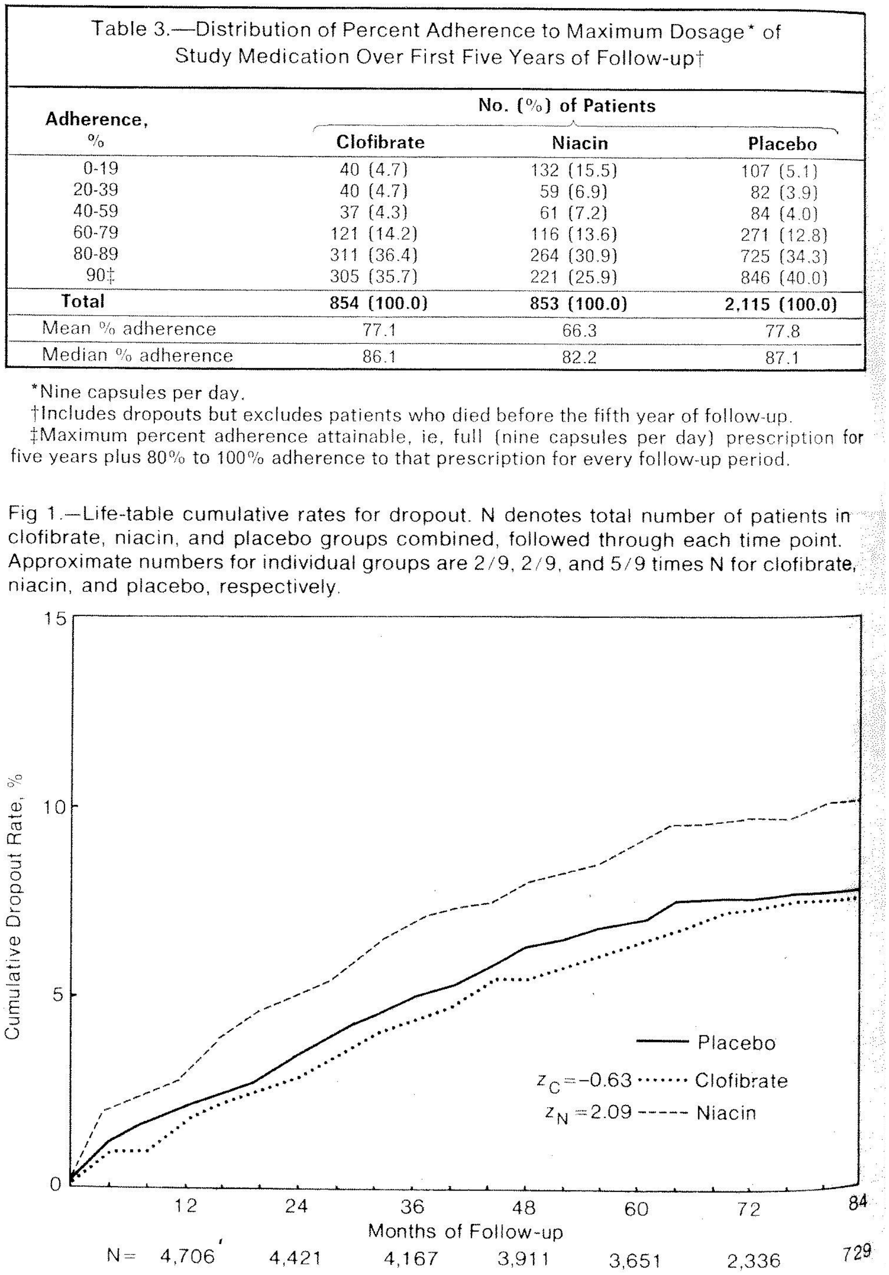

Table 3.—Distribution of Percent Adherence to Maximum Dosage* of Study Medication Over First Five Years of Follow-up†

Adherence, %	No. (%) of Patients		
	Clofibrate	Niacin	Placebo
0-19	40 (4.7)	132 (15.5)	107 (5.1)
20-39	40 (4.7)	59 (6.9)	82 (3.9)
40-59	37 (4.3)	61 (7.2)	84 (4.0)
60-79	121 (14.2)	116 (13.6)	271 (12.8)
80-89	311 (36.4)	264 (30.9)	725 (34.3)
90‡	305 (35.7)	221 (25.9)	846 (40.0)
Total	**854 (100.0)**	**853 (100.0)**	**2,115 (100.0)**
Mean % adherence	77.1	66.3	77.8
Median % adherence	86.1	82.2	87.1

*Nine capsules per day.

†Includes dropouts but excludes patients who died before the fifth year of follow-up.

‡Maximum percent adherence attainable, ie, full (nine capsules per day) prescription for five years plus 80% to 100% adherence to that prescription for every follow-up period.

Fig 1.—Life-table cumulative rates for dropout. N denotes total number of patients in clofibrate, niacin, and placebo groups combined, followed through each time point. Approximate numbers for individual groups are 2/9, 2/9, and 5/9 times N for clofibrate, niacin, and placebo, respectively.

FIGURE 5. The distribution of adherence and cumulative dropout rate for clofibrate, placebo and niacin. Reproduced without permission from [19].

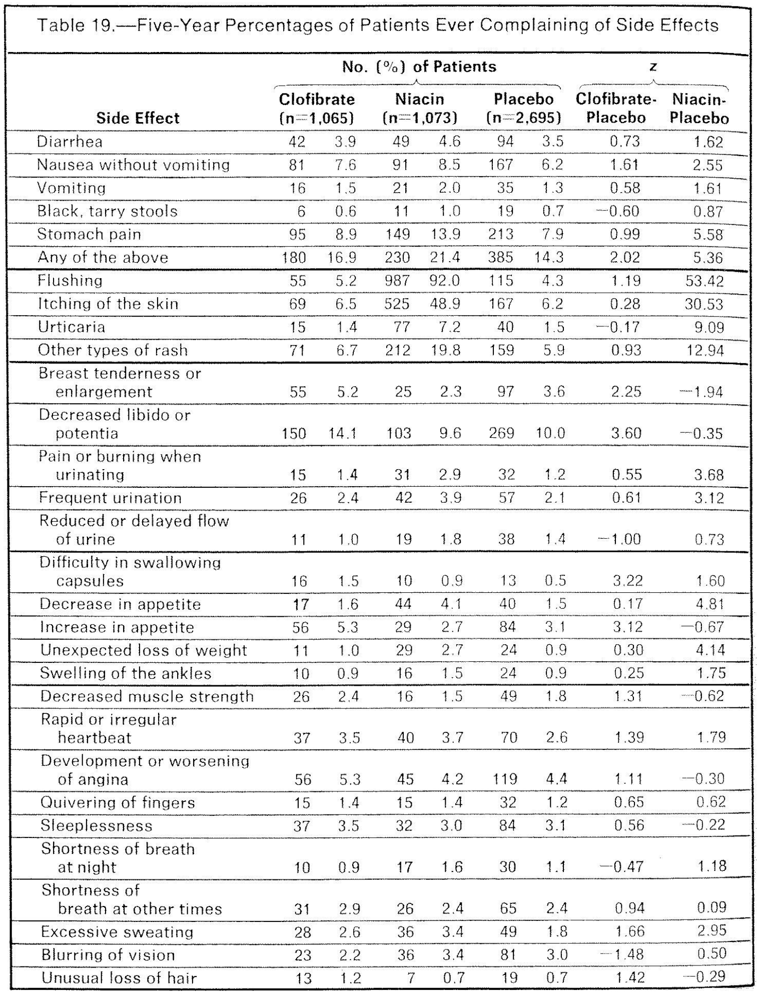

Table 19.—Five-Year Percentages of Patients Ever Complaining of Side Effects

	No. (%) of Patients						z	
Side Effect	Clofibrate (n=1,065)		Niacin (n=1,073)		Placebo (n=2,695)		Clofibrate-Placebo	Niacin-Placebo
Diarrhea	42	3.9	49	4.6	94	3.5	0.73	1.62
Nausea without vomiting	81	7.6	91	8.5	167	6.2	1.61	2.55
Vomiting	16	1.5	21	2.0	35	1.3	0.58	1.61
Black, tarry stools	6	0.6	11	1.0	19	0.7	−0.60	0.87
Stomach pain	95	8.9	149	13.9	213	7.9	0.99	5.58
Any of the above	180	16.9	230	21.4	385	14.3	2.02	5.36
Flushing	55	5.2	987	92.0	115	4.3	1.19	53.42
Itching of the skin	69	6.5	525	48.9	167	6.2	0.28	30.53
Urticaria	15	1.4	77	7.2	40	1.5	−0.17	9.09
Other types of rash	71	6.7	212	19.8	159	5.9	0.93	12.94
Breast tenderness or enlargement	55	5.2	25	2.3	97	3.6	2.25	−1.94
Decreased libido or potentia	150	14.1	103	9.6	269	10.0	3.60	−0.35
Pain or burning when urinating	15	1.4	31	2.9	32	1.2	0.55	3.68
Frequent urination	26	2.4	42	3.9	57	2.1	0.61	3.12
Reduced or delayed flow of urine	11	1.0	19	1.8	38	1.4	−1.00	0.73
Difficulty in swallowing capsules	16	1.5	10	0.9	13	0.5	3.22	1.60
Decrease in appetite	17	1.6	44	4.1	40	1.5	0.17	4.81
Increase in appetite	56	5.3	29	2.7	84	3.1	3.12	−0.67
Unexpected loss of weight	11	1.0	29	2.7	24	0.9	0.30	4.14
Swelling of the ankles	10	0.9	16	1.5	24	0.9	0.25	1.75
Decreased muscle strength	26	2.4	16	1.5	49	1.8	1.31	−0.62
Rapid or irregular heartbeat	37	3.5	40	3.7	70	2.6	1.39	1.79
Development or worsening of angina	56	5.3	45	4.2	119	4.4	1.11	−0.30
Quivering of fingers	15	1.4	15	1.4	32	1.2	0.65	0.62
Sleeplessness	37	3.5	32	3.0	84	3.1	0.56	−0.22
Shortness of breath at night	10	0.9	17	1.6	30	1.1	−0.47	1.18
Shortness of breath at other times	31	2.9	26	2.4	65	2.4	0.94	0.09
Excessive sweating	28	2.6	36	3.4	49	1.8	1.66	2.95
Blurring of vision	23	2.2	36	3.4	81	3.0	−1.48	0.50
Unusual loss of hair	13	1.2	7	0.7	19	0.7	1.42	−0.29

Figure 6. The distribution of side effects to clofibrate, placebo and niacin. Reproduced without permission from [19].

evaluation of outcomes. Even if the procedure is followed to the letter, the effect might not be achieved, if there is some other way for patients to discover their treatment assignment. For example, if the treatment causes a rash in 30% of patients (and the placebo group is rash-free), then 30% of the treatment group will be able to guess correctly that they're taking the treatment, even if no-one told them. That knowledge may influence how both patients and physicians assess their medical outcomes. For this reason, I define unblinding not in terms of whether trialists failed to follow the procedure, but rather in terms of whether blinding could have its intended effect.

It's also worth noting how unblinding is measured. Trialists typically (when they measure blinding at all) ask participants (and sometimes assessors) to guess which group they had been assigned to. They call the trial unblinded if the participants, and/or the physicians evaluating their outcomes, can guess the patients' group allocation better than chance. In other words, they check for independence or dependence between *Allocation* and $Beliefs$ (where $Beliefs$ is just shorthand for "beliefs about allocation").

According to the Causal Markov Condition, if two variables are unconditionally associated, there must be some causal connection between them – either a path from one to the other, or else a common cause of both. Now, *Allocation* is randomized, so it has no causes; $Beliefs$ cannot cause it, and neither can there be a common cause of *Allocation* and $Beliefs$. As a result, if we see an association between *Allocation* and $Beliefs$, it must be the result of a directed path from *Allocation* to $Beliefs$. This makes intuitive sense – a trial is unblinded when the group I'm *actually* assigned to influences what group I *think* I'm assigned to. I'll adopt this causal interpretation as my definition of unblinding.

> DEFINITION 1: *Unblinding.* A trial is unblinded iff there is a directed path from *Allocation* to the participants' $Beliefs$ about their allocation, and/or the $Beliefs$ of the physicians evaluating the patients' outcomes.

Given this formulation, we can talk about degree and mechanism of unblinding. If a trial is unblinded to a higher degree, the association between *Allocation* and $Beliefs$ will be stronger – i.e. more patients may be able to guess correctly, or the ones that guess correctly may be more certain about their guesses.[13]

We care about unblinding because we worry that in unblinded trials, *Allocation* might influence *Outcome* through reporting bias or differential treatment of the two groups, rather than through the intended treatment (see Section 2.1). So the mechanism of unblinding matters, and will be treatment- and trial-specific. Here's an example of a possible mechanism: I might get a rash that allows me (and my doctor) to infer I'm in the active treatment group. Or I might make the same inference because I experience a noticeable treatment benefit. Alternatively, if I am a relatively stable patient, I might notice that my condition has not changed at all since starting treatment, and infer that I have been given a placebo. The mechanism determines *which* patients become unblinded. Only patients who are susceptible to adverse effects, responsive to the treatment, or aware of their condition remaining stable (respectively) can be unblinded in these ways.

In an unblinded trial, we might represent reporting bias and differential treatment by adding a direct edge from *Allocation* to *Outcome*. We would also have to add an edge from *Allocation* to *Adherence*, because as soon as I know my allocation, that may influence whether I choose to adhere (for example, if I learn I'm taking a placebo, I may not be motivated to continue). Figure 7 represents this situation.

Unfortunately, Figure 7 doesn't capture the mechanism of unblinding we just discussed. In the example, I learn my *Allocation* by experiencing some actual *Outcome* caused by

[13] Note that participants might know their assignment but not know that they know it – if, for example, they do not feel confident in their best guess. In this case they might guess better than chance on a forced-choice questionnaire, but guess at chance levels if allowed to answer, 'I don't know'.

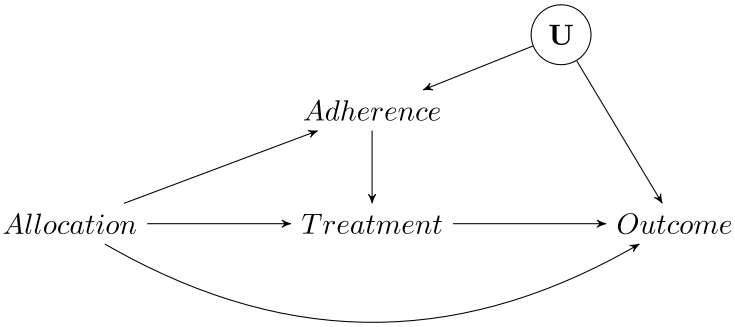

FIGURE 7. Causal structure of a Trial with Non-Adherence and Unblinding

Treatment. Barring cases where the two regimens are identifiable from the outset,[14] all instances of unblinding must work this way. So it seems like all the effects of *Allocation* should go through *Treatment* after all; we must remove the direct edges from *Allocation* to *Outcome* and *Adherence*. But then the nuisance effects of *Allocation* are included in the edge from *Treatment* to *Outcome*, which is exactly what we want to measure, so the graph does not represent the problem.

We want to capture both the nuisance pathways opened up by unblinding, and the intuition that in trials designed to be double-blind, unblinding always occurs as a consequence of some treatment effect. To do this we need to expand our graph to show the process as a time-series. Even if the initial round of treatment is perfectly blinded, the problem is that *subsequent* treatments become unblinded as a consequence of the first round of treatment.

7.1. **Time series representation of an unblinded trial.** As with all the graphical representations so far, we will construct the time-series graph using our background knowledge of the causal connections. The time-series representation is more complicated, however. In order to capture the way in which unblinding affects outcomes through a "nuisance pathway", we must add a new variable on that pathway: patient and doctor *Beliefs* (i.e. their beliefs about *Allocation*). In a blinded trial, *Beliefs* should be *d*-separated from *Allocation* given the empty set, whereas in an unblinded trial they should be *d*-connected.

Figure 8 represents this *d*-separation relationship, plus several key features of the causal process. To reduce graphical "spaghetti" I have represented *Outcome* as occurring at only one time (the end of the trial), and likewise with the set of unobserved common causes **U**, but in principle both could be made part of the time-series. I have not included an

[14] A trial of surgery vs. medication, or medication vs. counseling, or exercise vs. diet, will necessarily be unblinded like this, and thus might be represented by Figure 7. Trials where the regimens appear very similar – for example, a trial comparing two different laparoscopic surgical techniques, or two different QD pharmaceutical regimens – will not. However even among pharmaceutical trials, it is possible that participants could be unblinded immediately if, for example, the pills look, taste or smell like an over-the-counter preparation of the same drug. In the text I limit my discussion to trials where blinding was initially successful, so all edges out of *Allocation* must go through *Treatment*.

edge from *Allocation* to B_1, solely because I am limiting my discussion to trials that were intended to be blinded – where some effort was made to conceal allocation – so that the only way *Allocation* could affect $Beliefs$ would be through $Treatment$. The double-lined arrows from $Treatment_t$ to $Beliefs_{t'}$ represent the edges that are present in unblinded trials, but absent in blinded trials. For easy comparison, Figure 9 represents a blinded trial (without the double arrows).

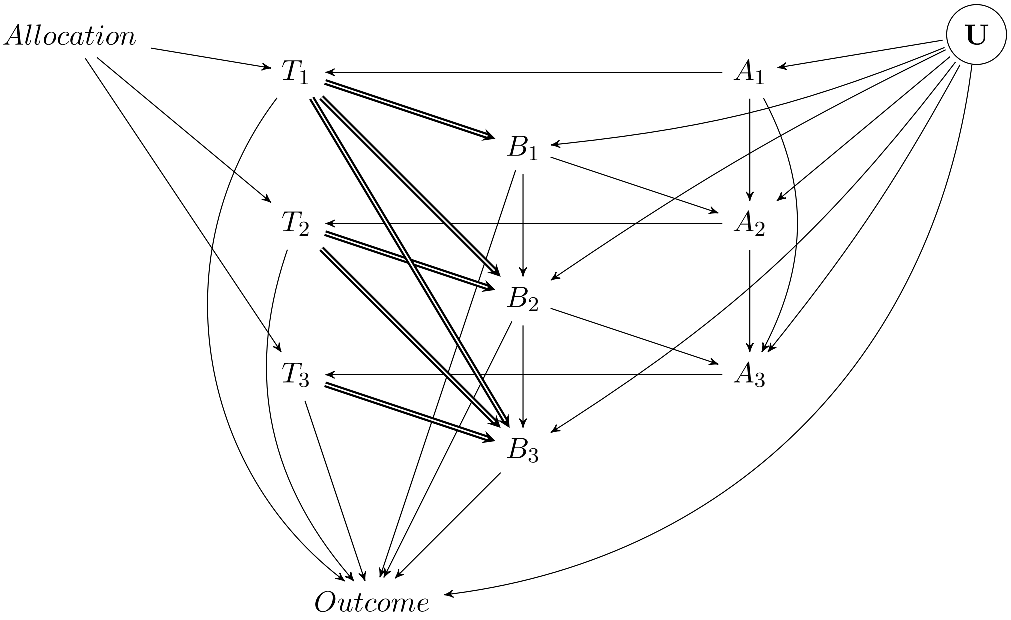

FIGURE 8. Time Series of a Trial with Non-Adherence (and double arrows for effects of Unblinding). A_t stands for adherence at time t, T_t for treatment received at t, B_t for patient beliefs at t.

8. Bias due to unblinding

Before we can determine whether our analysis is biased, we have to decide which effect we are interested in estimating. But now the picture is much more complicated; it is by no means clear what we wish to estimate. Previously we were interested in two effects: (a) the direct effect of *Treatment* on *Outcome*, and (b) the portion of the total effect of *Allocation* on *Outcome* that went through *Treatment*. The aim was to exclude the influence of expectation effects, attribution effects, differential treatment of the two groups, and all such phenomena that might, in an unblinded trial, produce a comparison that did not seem fair.

Now, however, we have to consider pathways like: $Allocation \to Treatment_t \to Beliefs_t \to Adherence_{t+1} \to Treatment_{t+1} \to Outcome$. This might represent a causal story like so: the drug appeared to have no effect, so I came to believe I was taking a placebo, so I

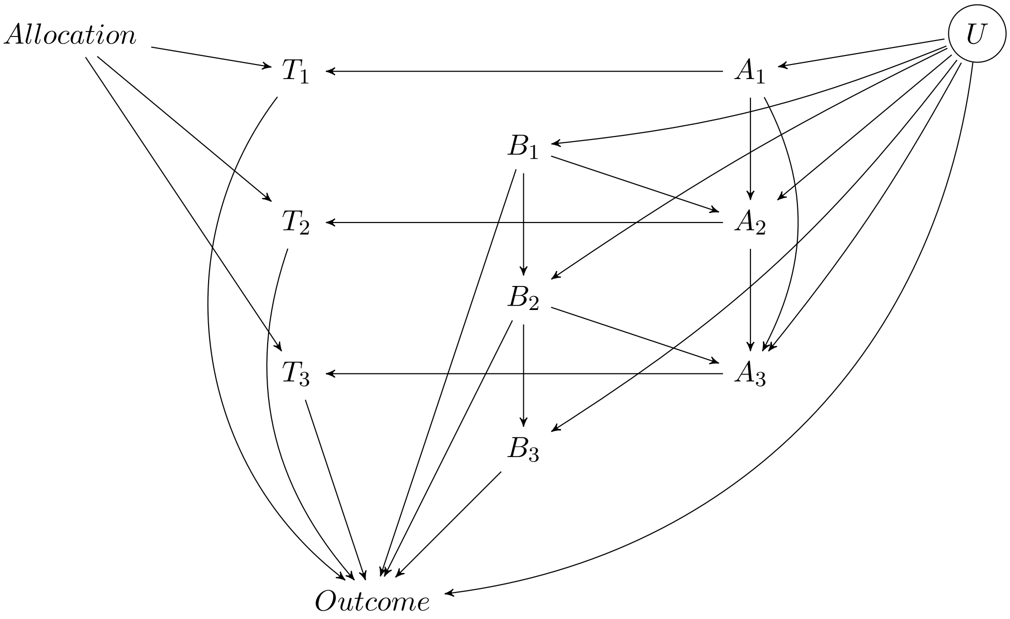

FIGURE 9. Time Series of a Trial with Non-Adherence (but no Unblinding). A_t stands for adherence at time t, T_t for treatment received at t, B_t for patient beliefs at t.

stopped adhering, which set my treatment level to zero, and this affected my outcome. Alternately, it might mean: the drug had noticeable side effects, so I came to believe I was taking treatment, which encouraged me to adhere, which raised my treatment level, affecting my medical outcome.

This pathway only exists in an unblinded trial, because it involves an edge from $Treatment$ to $Beliefs$. It may cause the two treatment groups to adhere to very different degrees (or medically different groups of people to adhere in each group), so it might seem to create an unfair comparison between groups, and on those grounds we might want to exclude it from influencing our analysis.

On the other hand, in this pathway, $Beliefs$ influences my *Outcome* indirectly – via my future $Treatment_{t+1}$ value – rather than directly, as with a reporting bias effect. And we *are* interested in the effects of that $Treatment_{t+1}$. We want to include every $Treatment_t \rightarrow Outcome$ edge in our effect of interest. Also, if one regimen is easier or more pleasant to follow, then that will increase its allocation effect (due to increased adherence), and we wish to measure this benefit when we calculate the total effect of prescribing the regimen.

By contrast, paths like $Allocation \rightarrow Treatment \rightarrow Beliefs \rightarrow Outcomes$ must be excluded, because they represent exactly those reporting biases, differential co-prescriptions, etc. that we wish to prevent from influencing our estimate.

We also want to exclude paths that contain a collider, because they represent statistical artifacts. Those pathways will only be open if we condition on the collider, or one of its descendants, in the analysis. They cannot represent causal effects of treatment.

Thus, even if I remain agnostic about whether to include paths like $Treatment_t \rightarrow Beliefs_t \rightarrow Adherence_{t+1} \rightarrow Treatment_{t+1}$, there are still three desiderata for what to include when we estimate the effect of *Allocation* on *Outcomes* and the effect of $Treatment$ on *Outcomes*. We should:

(1) Include every direct edge from $Treatment$ to *Outcomes*,
(2) Exclude every direct edge from $Beliefs$ to *Outcomes*, and
(3) Exclude every pathway that contains a collider.

These are easy to satisfy in a blinded trial (see Figure 9). To estimate the effect of *Allocation* on *Outcome*, it's sufficient to calculate the association between *Allocation* and *Outcome*; there is no confounding. To estimate the effect of $Treatment$ on *Outcome*, we can simply condition on all time-steps of *Adherence* (as in a Per Protocol analysis); this will block the confounding pathways through $\mathbf{U}$.

However, in the unblinded trial shown in Figure 8, our three desiderata cannot be jointly satisfied. Consider B_1. If we do not condition on it,[15] then the pathway $Allocation \rightarrow T_1 \rightarrow B_1 \rightarrow Outcome$ is active, violating desideratum (2). If we do condition on it, the pathway $Allocation \rightarrow T_1 \rightarrow B_1 \leftarrow \mathbf{U} \rightarrow Outcome$ becomes active, violating (3). Thus given the time-series representation above, we cannot satisfy desiderata (2) and (3) by conditioning.

This is true whether we are trying to estimate the Allocation Effect or the Treatment Effect. However, if we are trying to estimate the Treatment Effect, we have an additional worry. We must condition on *Adherence* to block the back-door pathway through $\mathbf{U}$; but this activates *Adherence* as a collider, and opens up the path $Treatment \rightarrow Beliefs \rightarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome$. We can break this pathway if we condition on $Beliefs$, but then we open up the pathway $Allocation \rightarrow Treatment \rightarrow Beliefs \leftarrow \mathbf{U} \rightarrow Outcome$. There is no escape: a collider path always may be active in a per protocol analysis of an unblinded trial.

Thus, for both of the causal effects we're interested in, bias is introduced if and only if the trial is unblinded. However there is a difference in which pathways bias the two effects. Assuming that we don't condition on $Beliefs$ (because as far as I know, no analyst has conditioned on $Beliefs$ in these circumstances), the Allocation Effect will only be biased by a direct effect of $Beliefs$ on *Outcome*, whereas the Treatment Effect will be biased by both that effect, and also the pathway $Treatment \rightarrow Beliefs \rightarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome$. So intention-to-treat and per protocol analyses will be biased in the same circumstances, but per protocol analyses may be biased to a greater degree. This difference might be the reason for the extra suspicion of per protocol analyses.

Note that if we do condition on $Beliefs$, then our estimates of the *Allocation* effect and the $Treatment$ effect are both confounded by the same pathway: $Treatment \rightarrow Beliefs \leftarrow$

[15] In this section I suggest conditioning on patients' and doctors' $Beliefs$ about allocation in order to better estimate the effect of treatment. As far as I know this has not been done.

$\mathbf{U} \to Outcome$. This is a good reason to try measuring $Beliefs$ in more trials, and see whether conditioning on $Beliefs$ produces consistent results for the *Allocation* and $Treatment$ effects.

Part 3. Possible approaches to deal with confounding by unblinding

In this sorry situation, there are at least four possible approaches we can take:

(1) Find some estimator that is not biased by the existence of these nuisance pathways – one that can control for bias using some operation besides conditioning (perhaps by making parametric assumptions);
(2) Expand the causal structure to include extra variables that we can condition on to block the nuisance pathways;
(3) Argue for a given trial that the magnitude of the bias is small; or
(4) Assume some weaker consequence of blinding that is sufficient to eliminate bias (perhaps with support from #3).

I know of no solutions that use option (1). I'll cover options (2)–(4) in turn.

9. Expanding the causal structure

9.1. **Blocking the path through U.** Looking back at Figure 8, clearly the confounding through $\mathbf{U}$ is the source of all our woes. Authors who notice an association between *Adherence* and *Outcome* in the placebo group often propose plausible-sounding candidates for some of the variables in U. [24] For example, Simpson et. al (2006) comments that "diet, exercise, regular follow-up with healthcare professionals, immunisations, screening, and use of other drugs," as well as depression, could contribute to the association between *Adherence* and *Outcome*. [45] John Urquhart proposed that adherence to effective non-trial medication could explain practically all of the association between adherence and outcomes in the Coronary Drug Project. [49] This hypothesis is particularly plausible because people who adhere to one medication are likely to adhere to others, [6] and because the participants in the CDP were prescribed many non-trial medications (some of which were effective) [18, p. I-29, Table 11], [19, p. 376, Table 20].

In most trials these variables are not measured. Co-prescriptions may be measured, but *adherence* to those co-prescriptions is what matters, and it is not measured. In theory, a trial could be designed with the secondary goal of measuring enough candidate members of $\mathbf{U}$ to block the pathway through $\mathbf{U}$.

Unfortunately if the treatment is effective, it is impossible to tell whether a set of candidates $\mathbf{C}$ block the pathway through $\mathbf{U}$. We can do it in the placebo group, but not the active treatment group, for the following reasons:

Say we are looking only at the placebo group (i.e. conditioning on $Allocation = 0$). There, we expect that $Treatment$ has no direct effect on *Outcome*, because the patients are not receiving any treatment. So all the $Treatment \to Outcome$ edges vanish. Then the only pathways d-connecting *Adherence* to *Outcome* go through $Beliefs$ or $\mathbf{U}$. Thus, if we condition on our set of candidates $\mathbf{C}$ and also $Beliefs$, then we will d-separate *Adherence* from *Outcome* if and only if **C** blocks the pathway through **U**. So in the placebo group it suffices to check whether $Adherence \perp\!\!\!\perp Outcome | \{\mathbf{C}, Beliefs\}$.[16]

In the treatment group, however, we expect $Treatment$ to affect *Outcome* directly. Thus, even if we condition on $Beliefs$, and **C** blocks the pathway through **U**, we still have the active path $Adherence \rightarrow Treatment \rightarrow Outcome$, *d*-connecting *Adherence* and *Outcome*. We cannot block this path by conditioning on $Treatment$, because *Adherence* and $Treatment$ are so closely related – in fact, in the treatment group they are identical – that if we condition on $Treatment$ there will be no variation left in *Adherence*.

This difference between the groups matters, because there may be an interaction between **U** and $Beliefs$ when they influence *Adherence*. If there's no such interaction, then we can assume that if **C** blocks the pathway through **U** in the placebo group, then it also does so in the treatment group. Unfortunately, we can only check indirect indicators of this interaction, such as:

(a) That the patients with zero adherence to treatment have similar outcomes to the patients with zero adherence to placebo; or

(b) That the difference in outcomes between good adherers to treatment and good adherers to placebo is consistent with the allocation effect measured in an ITT analysis.

9.2. **Using mediator variables and the Front Door Criterion.** An alternative approach is to use Judea Pearl's Front Door Criterion, [37] a technique usually applied to observational studies. It involves finding a *mediator* of the effect of $Treatment$ on *Outcome* – that is, a variable on the causal pathway between them. In general we can picture the drug mechanism of action as having this form: $Treatment \rightarrow Mediator \rightarrow Outcome$. Real examples might include $AZT \rightarrow Viral\ Load \rightarrow CD4\ count$ (for HIV infection), or $Bisphosphonates \rightarrow Bone\ Density \rightarrow Fractures$ (for osteoporosis). If we have a good mediator, then our causal graph should look like Figure 10.

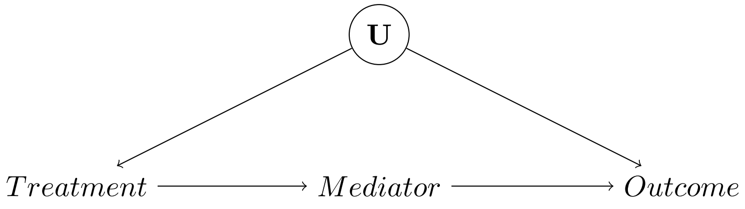

FIGURE 10. Mediator variable used for Front Door analysis

There is an open back-door pathway between $Treatment$ and *Outcome*, which cannot be blocked by conditioning, so it might look like the effect of $Treatment$ on *Outcome* is

[16] This is the test the CDP research group performed when they ran their multiple regression. They failed to break the association between *Adherence* and *Outcome*, indicating that they had not conditioned on the right variables. Their test was valid in both groups, because the treatment was completely ineffective. In trials of effective treatments it would only be a valid test in the placebo group – a point that the CDP research group did not make, and to the best of my knowledge, neither has anyone else who has repeated their argument.

hopelessly confounded. The key to the *Front Door* method is that the effect of $Treatment$ on $Mediator$ is unconfounded, and the effect of $Mediator$ on *Outcome* is unconfounded if we condition on $Treatment$ (blocking the back-door pathway through **U**). Thus, intuitively, we can estimate the causal effect of $Treatment$ on *Outcome* by estimating each step in the pathway, and then combining them. The Front-Door Formula (for discrete variables) is as follows:

$$P(O|do(T=t)) = \sum_m P(M=m|T=t) \sum_{t'} P(O=o|T=t', M=m)P(T=t')$$

where O represents *Outcome*, T is $Treatment$ and M is the $Mediator$. There is a natural generalization to continuous variables. The formula gives us the entire probability distribution of *Outcome* for any setting of $Treatment$, so for any comparison of two values of $Treatment$, we can use it to compute the average causal effect.

The problem is that there may be no *unconfounded* mediator variable. In our osteoporosis example, *Exercise* is probably a common cause of *Bone Density* and $Fractures$. If any unmeasured common causes fits the same structural position as *Exercise* or **V** in Figure 11, the Front Door method will fail. Likewise, if there is some a direct effect of $Treatment$ on *Outcome* that does not go through the $Mediator$, the Front Door method will fail to capture that part of the total effect of $Treatment$ on *Outcome*, even though it does capture the portion of the effect that goes through the $Mediator$. So the Front Door method may work in specific cases where we have very good background knowledge about the causal mechanism of the treatment, and believe we have an unconfounded mediator. Otherwise it will not be applicable.

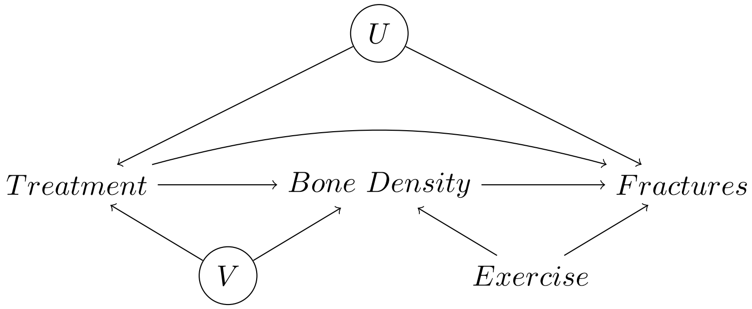

Figure 11. Causal structure of a Confounded Treatment Mechanism

10. Argue for a given trial that the magnitude of the bias is small

We might argue that the analysis of our particular trial is not biased. There are multiple ways to go about this. We could argue that the trial is blinded (or very close to blinded); in other words, we'd argue that the $Treatment$ → $Beliefs$ edges are absent, or (if our goal is a per protocol analysis) at least that the *Allocation* to *Adherence* path is absent.

Alternatively, we could argue that despite being unblinded, this has no effect (or almost no effect) on our results.

10.1. **Arguing that the trial is blinded.** If the trial is blinded, we expect *Allocation* to be unconditionally independent of $Beliefs$. Therefore, we can test the success of blinding empirically, assuming we have measured $Beliefs$.

In practice, however, trials rarely measure the success of blinding participants (and even fewer measure blinding of assessors). When they do, they often find that blind has been partly broken. [10] Furthermore, the methods for assessing the success of blinding are inconsistent and usually inadequate, as is the reporting of the measurements. [4] There exist proposals for better methods. [29] To reduce measurement error, a trial would need to measure blinding (a) in all the groups of people involved in assessing outcomes, prescribing cotreatment and influencing adherence, [8] (b) at multiple intervals throughout the trial, and (c) using at least a forced-choice format (possibly with the addition of a second format using an 'unknown' option, or an uncertainty scale).

If the trial is blinded, we would also expect *Allocation* to be unconditionally independent of *Adherence*.[17] This independence could also be tested empirically, and should be done whenever the researchers wish to condition on *Adherence* (as in a per protocol analysis).

Researchers could also argue for the likely success of blinding on mechanistic grounds. If the effect of treatment is not apparent to the participants (for example, treatments for hypertension) and the treatment has no noticeable side effects (or is being compared with an active placebo with a similar side-effect profile), one could argue that there is no way that participants could learn their allocation. Of course, this could only supplement rather than supplant the independence tests; there is no point arguing that participants could not learn their allocation if, in fact, $Beliefs$ are strongly correlated with *Allocation*. For historical trials, in which $Beliefs$ were not measured, seeing similar distributions of *Adherence*, *Side Effects*, *Dropout* and noticeable *Outcome* across treatment groups can provide indirect evidence of blinding.

10.2. **Arguing that although the trial is unblinded, this did not affect the results.** Alternatively, if our tests show that the trial was somewhat unblinded, we may still be able to argue that this does not affect the results of our analysis. In fact, researchers are put in this position whenever they measure participants' beliefs and find that the blind did not hold during the trial.

These arguments are about the strength of various mechanisms, which will naturally differ from trial to trial. The bias induced by conditioning on *Adherence* depends on the degree of confounding through **U**, which varies from trial to trial. [45] The direct effect of $Beliefs$ on *Outcome* can vary: Trials in which participants have access to powerful non-trial remedies afford greater scope for differential treatment. Reporting biases are stronger for "subjective" or "soft" endpoints, such as pain, than "objective" or "hard" ones, such as death. [41]

[17] In fact this would suffice for a per protocol analysis to be no more biased than an intention-to-treat analysis, even if the unblinding creates a directed path $Allocation \rightarrow Treatment \rightarrow Beliefs \rightarrow Outcome$.

Note, however, that apparently "hard" endpoints may be less objective than they seem. The Anturane Reinfaction Trial [16, 17] is an illustrative example. The trialists excluded 'non-analyzable' deaths; the primary endpoint was defined not as death from all causes, but *death from myocardial infarction (MI)*, so several deaths were deemed 'non-analyzable' because the autopsy found that they were due to sudden death rather than MI. Distinguishing MI from sudden death was difficult; a clot could be missed on autopsy and the death misclassified. Thus, what counts as a "hard" outcome can depend on the diagnostic/evaluative procedure and the comparison class of outcomes. The FDA performed an investigation of the trial and found that the vast majority of deaths excluded from analysis had been in the Anturane group rather than the placebo group, enough to nullify the trial's findings. [48]

It is now impossible to know whether the Anturane researchers really were unblinded, or if they were just astoundingly unlucky (as they claimed [44]). The point is that it would be possible for unblinded researchers to selectively report an endpoint as hard as 'death from myocardial infarction'.

10.3. **Greenland's argument about colliders vs. common causes.** Sander Greenland has argued that under several different parameterizations, conditioning on colliders introduces less bias than failing to condition on common causes. [14] This has the implication that if one variable is both a collider (one one pathway) and a common cause (on another), we are better off conditioning on it than not doing so. In the case of clinical trials, *Adherence* is not a common cause on the path $Treatment \leftarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome$, but it does block a path through the set of common causes $\mathbf{U}$, whereas it is a collider on the path $Allocation \rightarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome$. As such, conditioning on *Adherence* may introduce less bias of the $Treatment \rightarrow Outcome$ effect than it eliminates. However, this is cold comfort to the biostatistics community, which sets extremely conservative standards for inference in randomized controlled trials (as exemplified by their interpretation of the CDP).

11. Assume some consequence of blinding.

By far the most common approach to analysis is to assume some consequence of perfect blinding that is sufficient to eliminate bias, but is weaker than assuming *all* the consequences of blinding. This is obviously no guarantee against bias; the result of the analyses hold conditional on the assumptions made.

I'll cover three major approaches to analysis, all of which assume a consequence of blinding:

(1) Intention to treat (ITT)
(2) Per protocol
(3) Instrumental variables (IV)

11.1. **Intention To Treat.** Intent-To-Treat (ITT) analysis is so called because of who is included in the set of patients analyzed: namely, all the patients you *intended* to treat, regardless of whether they actually took the treatment. Besides specifying who is to be included, and the fact that the two treatment arms are compared, there are no constraints on what constituted an ITT analysis. It might be a simple t-test, a survival analysis, or anything else. Intent-to-treat is sometimes called a "principle" rather than an analysis because it can be applied to many different kinds of analyses. [34] The results of the analysis are supposed to represent the effect of intending to treat someone – that is to say, the effect of *Allocation* on *Outcome*, rather than the effect of $Treatment$. This typically produces an underestimate of both treatment benefits and adverse effects in patients who were actually treated.

Running an ITT analyses is less simple than it sounds. If participants drop out of the study and cannot be contacted, their final data cannot be included in the analysis (although data up until the point at which they dropped out may be included, for example in a survival analysis). This matters in an unblinded trial, because participants can drop out of the two groups for different reasons. The variable *Dropout* is in a similar structural position to *Adherence* – it is plausibly a child of both $Beliefs$ and $\mathbf{U}$. Yet unlike *Adherence*, *Dropout* is a selection variable; we are forced to condition on it. When we look at all our remaining participants, we are looking at a subset of our original sample in which $Dropout = 0$ for everyone in that subsample. As a result, if the trial is unblinded, then even in an ITT analysis, there is an open path that might bias our estimate of the Allocation Effect: $Allocation \rightarrow Treatment \rightarrow Beliefs \rightarrow Dropout \leftarrow \mathbf{U} \rightarrow Outcome.$[18]

One approach to dropout is to "impute" (i.e. make up) data for the missing participants, but this requires dubious distributional assumptions about how similar the missing participants are to the continuing ones, and in the worse case it would not only preserve but exacerbate any bias produced by participants self-selecting themselves out of the study for different reasons in the two arms. More appropriate is a sensitivity analysis, in which the analyst calculates the results conditional on various different outcomes for the missing participants, and takes the most extreme results to be bounds on the effect (although if there is substantial dropout the bounds of a sensitivity analysis may be uninformative). The rate of dropout limits how much it can affect our results, but so does the rate of outcome events. If the trial is supposed to prevent a rare event, such as fractures in patients with osteoporosis, even a small proportion of dropouts could bias the results substantially, so long as it is not small relative to the number of fractures.

Another option is Last Observation Carried Forward (LOCF) analysis, which assumes that participants who drop out will tend to stay in the same medical state as when they dropped out; this can be biased if the sample tends to deteriorate or improve (e.g. Alzheimers, or the common cold) and more participants drop from one group than the other, so a correction for this effect and a sensitivity analysis are also useful. A last option, widely used in economics to control for selection bias, is to apply the Heckman correction, although that also depends on parametric assumptions.

ITT relies on the assumption of blinding for two things. Firstly, blinding implies that there is no differential treatment of the two groups and no reporting bias. Thus in the static

[18] Sometimes researchers ignore this problem and simply analyze the data they have available. Although this is often mislabeled ITT, it should properly be called an 'available case' analysis. [23]

causal graph – see Figure 3 for a blinded version and Figure 4 for an unblinded one – ITT assumes no direct effect of *Allocation* on *Outcome* that does not go through $Treatment$. Secondly, blinding implies that no bias is induced by $Dropout$. However, many forms of ITT analysis weaken this to something like, "despite potential bias induced by dropout, the true result is within the bounds of our sensitivity analysis", or "our LOCF analysis is not biased by differential dropout".

ITT tells us the effect of assigning a patient to take the treatment. What about the effect of taking it?

11.2. **Per Protocol.** If we are interested in the effect of *taking* the treatment, rather than just being assigned the treatment, the first thing we might try is a *per protocol* analysis, so called because it aims to measure the effect of following the protocol. Traditionally a per protocol analysis means an analysis that excluded non-adherent participants (by contrast to an ITT analysis, which includes them).

However, there is no reason a per protocol analysis could not be made more sophisticated. The essential idea is that we are conditioning on *Adherence*. If *Adherence* is not binary but multi-valued discrete, we could compare the two groups stratified by adherence level; if it is continuous, we could regress *Outcome* onto *Adherence* and compare the regression slopes in the two groups. This would give us something similar to a dose-response curve.[19] We could even use non-parametric regression if we do not want to assume a particular parametric relationship, and compare the areas under the regression curve for each group.

One variation of this approach was pursued by Efron & Feldman (1991). In order to estimate the dose-response curve for cholestyramine, they regressed *Outcome* onto *Adherence Quantile* and compared the slopes between the two groups. Efron and Feldman did not assume a linear dose-response relationship, opting instead for a quadratic model. [9][20]

Per protocol analyses rely on the assumption of blinding for the same reasons ITT analyses do: they assume there is no reporting bias and no differential treatment of the two groups, and that there is no bias induced by conditioning on *Dropout*. However, per protocol analyses need a third consequence of blinding: that there is no bias induced by conditioning on *Adherence*. Returning to the time-series graph of an unblinded trial (reprinted here as Figure 12), it's clear that conditioning on *Adherence* opens up pathways of the form $Allocation \rightarrow Treatment \rightarrow Beliefs \rightarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome.$

Strangely, the Cochrane Handbook instructs authors to perform a sensitivity analysis for bias induced by *Dropout*, but gives no analogous instruction for bias induced by conditioning on *Adherence*, saying instead that intent-to-treat analyses should always be preferred

[19] Because the level of *Adherence* is not randomized, to interpret this as a true dose-response curve, we have to assume that the effect of a given dose is independent of adherence-related patient characteristics.

[20] Note that Efron and Feldman did not refer to their analysis as 'per protocol', and it is much more sophisticated than the typical analysis that is labeled 'per protocol'. Also, the entire analysis was framed as a potential outcomes analysis. Lastly, the choice to regress on *Adherence Quantile* instead of *Adherence* was made because the distributions of *Adherence* were different in the two treatment groups, which implies an *Allocation* → *Adherence* edge. Efron and Feldman assumed that patients in the same quantiles in each group were comparable, which amounts to assuming that *Allocation* to treatment just decreases *Adherence* by some constant fraction, rather than interacting with **U**.

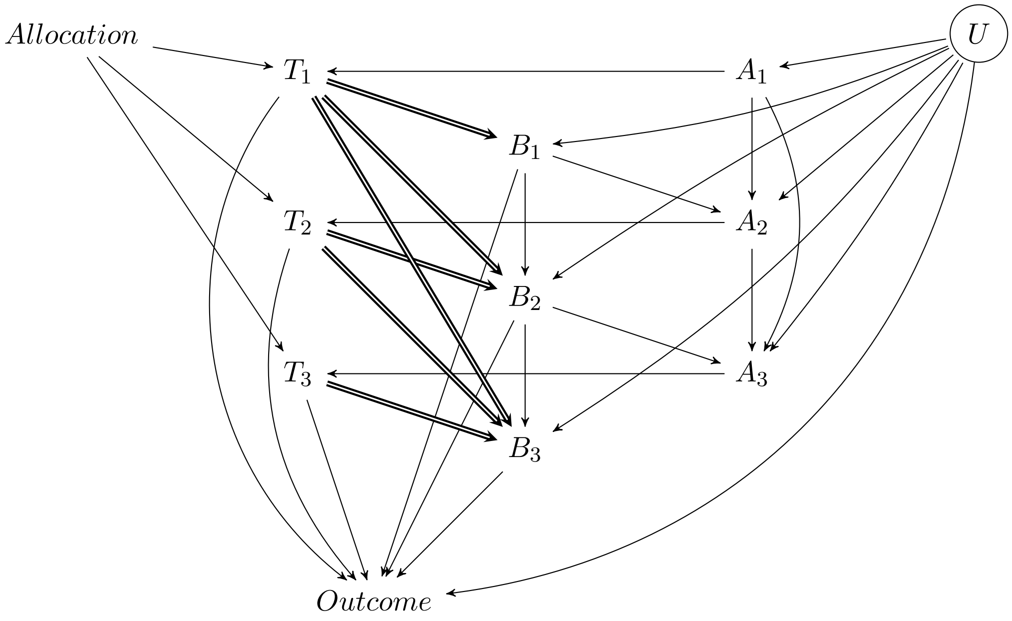

FIGURE 12. Time Series of a Trial with Non-Adherence (and double arrows for effects of Unblinding). A_t stands for adherence at time t, T_t for treatment received at t, B_t for patient beliefs at t.

to per protocol. However, in theory, we could perform a sensitivity analysis to assess the risk of bias induced by *Adherence*. In fact, because we have complete outcome data on the non-adherers (unlike the participants who drop out), the risk of bias can be assessed in more detail.

11.3. **Instrumental Variables.** Instrumental Variables (IV) analyses are a family of techniques that use a combination of graphical properties and additional parametric assumptions to either identify, or bound, the causal effect of *Treatment* on *Outcome*. The graphical structure assumed by IV methods is represented by Figure 13. Here, *Allocation* serves as an *instrument*. To qualify as an instrument, *Allocation* must affect *Outcome* only

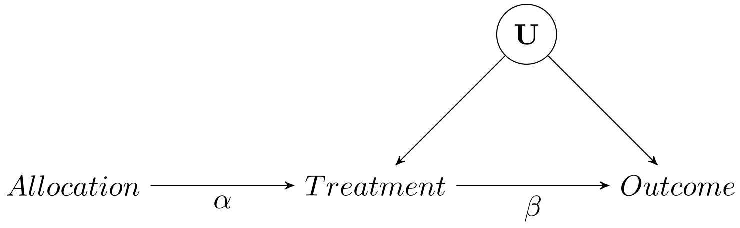

FIGURE 13. IV methods assume a randomised, double-blind trial with non-adherence

through $Treatment$, not directly – so IV methods must assume a blinded trial, to exclude direct effects of *Allocation* via reporting biases and differential treatment. Also, there must be no common cause of *Allocation* and *Outcome* (which is guaranteed by randomization). IV methods, like ITT and per protocol, also assume no bias induced by *Dropout*.

The IV approach essentially uses the fact that the effects of *Allocation* on $Treatment$, and of *Allocation* on *Outcome*, are unconfounded; the effect of $Treatment$ on *Outcome* should then be calculable from those two effects. Unfortunately, unlike in the case of the Front Door technique (see Section 9.2), to identify the causal effect using instrumental variables we also need to make some parametric assumptions. Hernàn and Robins (2006) demonstrate that it suffices to assume a linear or log-linear model, or a 'monotonic' model, where monotonic in this case means that $Treatment$ is a monotonic function of *Allocation*. In other words, if you would take the treatment given you'd been assigned to placebo, then you would also take the treatment if you were assigned to treatment. However, none of these parametric assumptions are testable from the data, and all three produce different estimates of the treatment effect, so it may be unclear which to use (if any). [21]

In a linear model, the IV method aims to estimate the linear coefficient β (see Figure 13). Because the effect of *Allocation* on $Treatment$ is unconfounded, the coefficient α can be estimated using the observed covariance: $Cov(Allocation,\ Treatment) = \alpha Var(Allocation)$. Likewise, the effect of *Allocation* on *outcome* can be estimated as, $Cov(Allocation,\ Outcome) = \alpha\beta Var(Allocation)$. Then the IV estimator is just the ratio of the two effects:

$$\begin{aligned}\beta &= \frac{Cov(Allocation,\ Outcome)}{Cov(Allocation,\ Treatment)} \\ &= \frac{E[Outcome|Allocation = 1] - E[Outcome|Allocation = 0]}{E[Treatment|Allocation = 1] - E[Treatment|Allocation = 0]}\end{aligned}$$

Where the second equality holds because *Allocation* is binary. Note, however, that when the denominator of the fraction is small – as in the case where adherence is very low – the estimator will be very high variance, and if *Allocation* is not a valid instrument (as in an unblinded trial) the bias introduced will be magnified.

11.3.1. *Bounding the Average Causal Effect using finite response variables.* Pearl extends IV methods to compute nonparametric bounds on the Average Causal Effect (ACE) of $Treatment$ on *Outcome*. [37, pages 262–9]

The technique relies on dichotomizing adherence and outcome, and then taking advantage of the limited number of possible ways that the unobserved variable U could influence adherence and outcome. Say we assume the graph in Figure 14. We would ideally like to condition on U, but unfortunately U is unobserved.

So instead, we posit two extra unobserved variables, R_T and R_O, shown in Figure 15. These don't have a mechanistic interpretation; they are just a mathematical fiction we use to represent all possible combinations of ways that U could influence $Treatment$ and *Outcome*, respectively. Pearl calls them "finite response variables".

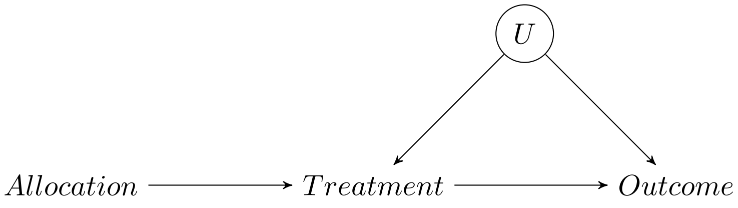

FIGURE 14. The causal structure of a Trial with Non-Adherence

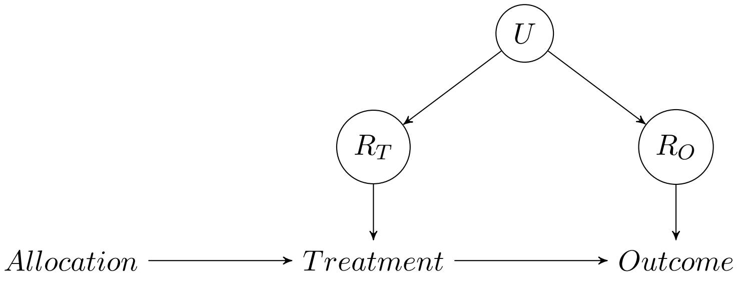

FIGURE 15. Showing finite response variables

Note that *Treatment* is binary, and it has one other binary parent, *Allocation*. Thus there are only four possible effects that R_T could have on *Treatment*: (1) make $Treatment = 0$, regardless of the value of *Allocation*; (2) make $Treatment = 1$, regardless of the value of *Allocation*; (3) if $Allocation = 0$, make $Treatment = 0$, and if $Allocation = 1$, make $Treatment = 1$; and (4) if $Allocation = 0$, make $Treatment = 1$, and if $Allocation = 1$, make $Treatment = 0$. These four "effects" of R_T can be mapped to four states. The states have an intuitive interpretation as a person's tendency to adhere to medication, and have been named 'never-taker', 'always-taker', 'complier' and 'defier', respectively. The probability that R_T takes on one of these four values can be thought of as the probability that an individual falls into each of these four categories of adherence behavior.[21] *Treatment* is then a deterministic function of *Allocation* and R_T – for example, if $Allocation = 1$ and $R_T = complier$, then $Treatment = 1$, and so on.

The situation is analogous with R_O and *Outcome*, because *Outcome* is also binary, and has one other binary parent, *Adherer*. The effects of R_O are then 'never-improve', 'always-improve', 'helped' and 'harmed' by treatment. *Outcome* becomes a deterministic function of *Treatment* and R_O. Then the value of *Outcome* given we intervene on *Treatment*

[21] Note that because R_T and R_O are mathematical fictions, whether someone counts as (say) a 'complier' or a 'defier' can change from trial to trial, and should not be interpreted as an enduring trait. The value of R_T is a counterfactual statement that only holds for this particular trial.

becomes:

$$P(Out = 1|do(Treat = 1)) = P(R_O = helped) + P(R_O = always\ improve)$$
$$P(Out = 1|do(Treat = 0)) = P(R_O = hurt) + P(R_O = always\ improve)$$

So the average causal effect of assigning $Treatment$ to 1 instead of 0 is:

$$P(Out = 1|do(Treat = 1)) - P(Out = 1|do(Treat = 0)) = P(R_O = helped) - P(R_O = hurt)$$

We can then express the probabilities of the observed variables as sums of the joint probabilities of the unobserved R_T and R_O. For example:

$$\begin{aligned} &P(Outcome = 1, Treatment = 0|Allocation = 1) = \\ &\qquad P(R_T = never\ taker, R_O = hurt) + P(R_T = never\ taker, R_O = always\ recover) \\ &\qquad\quad + P(R_T = defier, R_O = hurt) + P(R_T = defier, R_O = always\ recover) \end{aligned}$$

Likewise, the Average Causal Effect can be written out as a sum of the probabilities of the unobserved R_T and R_O. Then to find upper and lower bounds on the effect, we must first maximize this sum, and then minimize it, subject to the constraints. Every expression of the observed probabilities as a function of the unobserved ones functions as a constraint, as does the requirement that all probabilities sum to 1. Happily, these are all linear constraints, so the maximization and minimization tasks can be solved by linear programming. The problem is small enough that it can be solved analytically, producing precise bounds on the average causal effect.

The finite response variable approach requires the same graphical structure as a parametric IV analysis, so it also assumes a blinded trial, with no direct effect of *Allocation* on *Outcome*. The great strength of this approach is that it makes no parametric assumptions; it requires only that all the observed variables be dichotomous.

However, although every continuous variable can be coarsened into a dichotomous one, doing so does not always preserve the *d*-separation relationships satisfied by the original variable. Coarsening can introduce measurement error. Because the finite response variables approach requires the *d*-separation relationships in Figure 15, the approach effectively assumes that all of the information in $Treatment$ and $Outcome$ can be captured by binary variables. This may fail if, for example, we have an inverted U-shaped dose-response relationship.

11.4. **Similarities and differences in the causal assumptions of these methods.** Intent-to-treat, per protocol, parametric instrumental variables, and instrumental variables with finite response variables all have one thing in common: they all have to assume the trial was blinded. However, they don't all need this assumption for the same reasons. In particular, three structural consequences of blinding can be assumed independently of one another:

1. Assume there is no bias introduced through the directed path $Allocation \rightarrow Treatment \rightarrow Beliefs \rightarrow Outcomes$, and
2. Assume that if we condition on *Adherence*, there is no bias introduced through the collider path $Treatment \rightarrow Beliefs \rightarrow Adherence \leftarrow \mathbf{U} \rightarrow Outcome.$

(3) Assume that there is no bias induced by the selection variable *Dropout*, on the path $Treatment \rightarrow Beliefs \rightarrow Dropout \leftarrow \mathbf{U} \rightarrow Outcome$.

Obviously (1), (2) and (3) will all be true in a blinded trial, but if the trial is unblinded, the strength of the pathways matters. The bias introduced through the first path might be relatively weak while the second path was relatively strong, making (1) a reasonable assumption while (2) was not.

In fact, intention To Treat (ITT), Instrumental Variable (IV), and Average Causal Effect (ACE) analyses all require assumptions (1) and (3), but they do not require (2) because they don't involve conditioning on *Adherence*. Per protocol analyses require (1), (2) and (3), so they make strictly stronger causal assumptions. The FDA's and the Cochrane Collaboration's preference for ITT over per protocol analyses might be justified if they thought that assumption (1) and (3) were plausible, but (2) was not. However, (3) is structurally similar to (2), and (2) can be assessed in at least as much detail as (3) because we have data for the non-adherent participants, but not those who drop out. There will certainly be cases where, in a particular trial, the strength of the causal pathways makes (3) plausible but (2) implausible. However, there may equally well be cases where the reverse is true.

In addition to these causal assumptions, IV methods make parametric assumptions: they require linearity, or log-linearity, or monotonicity. When using finite response variables, it is possible to drop the parametric constraints, at the cost of assuming that we can dichotomize *Treatment* and *Outcome* without introducing measurement error. In the next section I will discuss relaxing the assumption of no measurement error, particularly for front-door, instrumental variable, and per protocol analyses.

Part 4. **Measurement error**

In all of the previous analyses, we assumed that we had measured all variables perfectly. By "measured perfectly" I mean we (a) captured the exact value of the variables, rather than some noisy approximation to those values, and (b) captured all the relevant information in each of the variables. There are several situations in which we might think our measurement gives us true values, but does not capture all the information. For example, we might be sampling a time-varying phenomenon, like blood glucose, at a slow rate – say once a month. Variations in blood glucose will occur continuously; massive (but brief) spikes in glucose could trigger a diabetic coma. But because we only measure once a month, we might not see the spike, even though we see the coma. Alternatively, we might take a

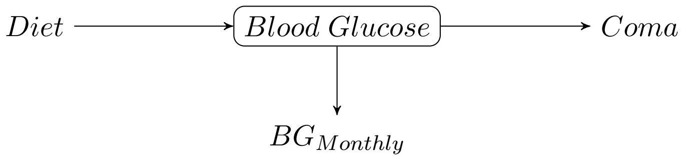

FIGURE 16. Sparse sampling of *Blood Glucose*

rough-grained partition of some variable when we really need a finer-grained partition. For example, we might dichotomise adherence into "good" and "poor", when in fact the precise number of doses taken, or the variation in dose timing, is relevant to, say, viral resistance. (Note that because $Treatment$ is a deterministic function of *Adherence* and *Allocation*, this would also give us a binary measure of $Treatment$.) Furthermore, our variable might not be what we are actually interested in. In the osteoporosis example, we measure *Bone Density* as a surrogate for *Bone Strength*. Not only does this introduce noise, it can also be misleading; for example, if calcium and magnesium are deposited in the outer sheath of the femur far more than the lattice of bone within, the bone can become very dense but also very brittle. Lastly, there may be simple noise in our measuring instruments. All these kinds of imperfect measurement can be called "measurement error", and clearly they occur very, very frequently.

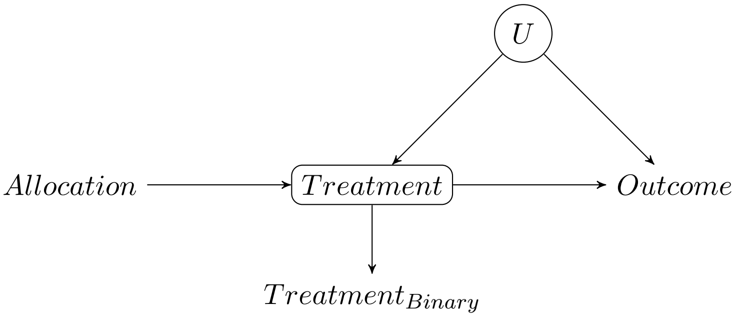

Figure 17. A rough-grained partition of $Treatment$

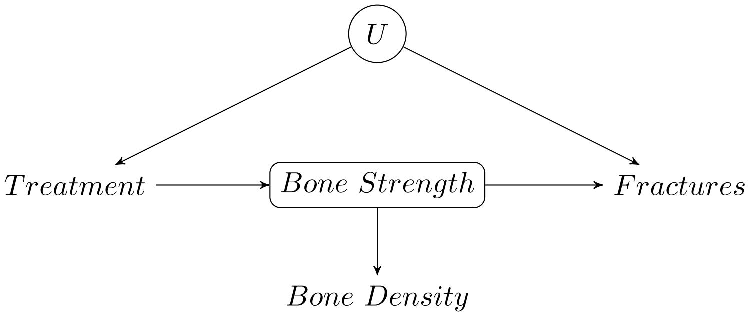

Figure 18. *Bone Density* as a surrogate for *Bone Strength*

All kinds of measurement error have a common graphical representation. The variable of interest is unobserved (shown circled in the graph) and we observe its child instead. It's clear from the representation that measurement error can cause huge problems for analyses that rely on d-separation relationships. If the error-prone variable is in the conditioning set, then we are not conditioning on it, but instead on its child, which is off the path of interest. For example, in the osteoporosis example, I would like to condition on the mediating variable, *Bone Strength*, to use the Front Door criterion. Unfortunately, conditioning on the measured variable, *Bone Density*, leaves open a direct path from $Treatment$ to $Fractures$ through *Bone Strength.* There's nothing I can do to block this path, so the Front Door Criterion simply won't work. There is no clever trick of conditioning that will save our analysis.

12. Linear models

However, if we assume the model is linear, we can do much more. *Linearity* is a parametric assumption; it says nothing about the causal structure of the graph, but puts strong constraints on the functional relationships between variables. A model is linear if every variable is a linear function of its parents (i.e. for any variable X, and its set of parents $P_1 ... P_n$, $X = \sum_i a_i P_i + \varepsilon_X$, where the a_i are real-valued coefficients and ε_X is a noise term independent of all the other variables.

Linearity is a very powerful assumption. It implies that for any unit change in the P_1, the value of X changes by some constant amount, a_1, regardless of the starting value of X. It implies that there is no interaction between the parents of X; they all influence X independently of each other. Linear models are very mathematically convenient, but in many cases strict linearity will be implausible.

However, often similar but weaker assumptions, like monotonicity, *are* plausible in a medical setting. If we know enough about the biochemical mechanisms in play, or if we have seen the results of a dose-response study, we might feel comfortable assuming that the treatment either increases bone strength or makes no difference, depending on the dose, but there is no dose at which it *decreases* bone strength. In these cases, we may perform simulation studies to see how far our results could go wrong if we assume a linear model when in fact the generating process is monotonic but not linear.[22] Monotonic models do not have the same nice theoretical properties as linear ones but in a host of real cases they may give similar results. Thus, it is worth looking at the properties of linear models even if we think linearity is implausible.

12.1. The Trek Rule.

Linear models are useful because they give us the *Trek Rule.* [47] The Trek Rule relates the linear coefficients of the edges in the graph to the covariance matrix of the variables, so we can learn about those edge coefficients from the covariances.

Definition 2: A *trek* is a path between two variables, X and Y, with the following features:

(1) There is a source node, Z, in the path (X, Y, or some variable in between them may be the source node).
(2) Every edge between X and Z is directed towards X
(3) Every edge between Z and Y is directed towards Y

[22] A good direction for future research.

So a directed path counts as a trek, and so does a path between two descendants through their common ancestor. Intuitively a trek has two "sides" – one from X to Z, one from Z to Y – and in the case of a directed path, one side of the trek is empty.

> Definition 3: *Trek coefficient*, a.k.a. *trek product.* In a linear model, for a given trek T, its *trek coefficient* is the product of all the coefficients of the edges on the trek, multiplied by the variance of the source of the trek.

Note that if the model has been standardised, so that all variables have mean zero and variance one, then the trek coefficient is just the product of the edge coefficients.

> Definition 4: The *Trek Rule*: In a linear model, the covariance between two variables X and Y is the sum of the trek coefficients for all treks between X and Y.

12.2. **Applying the Trek Rule to the Front Door with measurement error.** So say we have the following graph shown in Figure 19, where the model is linear and all variables have been standardised. This is the same as the Front Door case with measurement error of the mediating variable; I have just renamed the variables to keep the equations concise. So we want to learn the trek product $\alpha\beta$, as this corresponds to the effect of $Treatment$ on $Outcome$. Using the trek rule, we can write down the following formulae for the observed covariances:

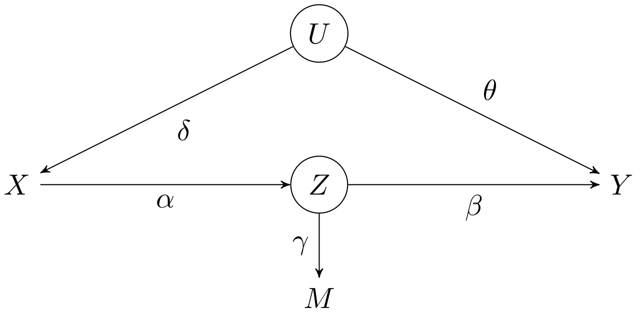

Figure 19. Front Door graph with measurement error of the mediating variable

$$Cov(X,Y) = \alpha\beta + \delta\theta$$
$$Cov(X,M) = \alpha\gamma$$
$$Cov(Y,M) = \beta\gamma + \gamma\alpha\delta\theta$$

Even though $\delta\theta$ can be treated as one parameter (because they always occur together), we still have far more unknowns than equations. We cannot express $\alpha\beta$ in terms of the observed covariances, so there's no way to identify the effect we're interested in from the covariance matrix.[23]

[23] In this section I am determining whether an effect is identifiable by writing out the constraints and then, if the effect is identifiable, doing algebra by hand to show how it is identified. For a general algorithm

However, if we had not one, but two different measurements of the mediating variable, as shown in Figure 20, the situation would be more hopeful. We would have three more observed covariances, giving us the following set of constraints:

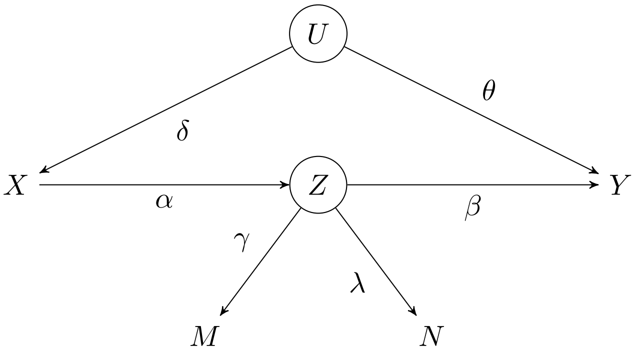

FIGURE 20. Front Door graph with two measurements of the mediating variable

$$\begin{aligned} Cov(X, Y) &= \alpha\beta + \delta\theta \\ Cov(X, M) &= \alpha\gamma \\ Cov(Y, M) &= \beta\gamma + \gamma\alpha\delta\theta \\ Cov(X, N) &= \alpha\lambda \\ Cov(Y, N) &= \beta\lambda + \lambda\alpha\delta\theta \\ Cov(M, N) &= \gamma\lambda \end{aligned}$$

Using $Cov(X, M)$, $Cov(X, N)$ and $Cov(M, N)$ we can identify the magnitude (though not the sign) of α, γ and λ, like so:

$$\begin{aligned} |\alpha| &= \sqrt{\frac{Cov(X, M)Cov(X, N)}{Cov(M, N)}} = \sqrt{\frac{\alpha\gamma\alpha\lambda}{\gamma\lambda}} \\ |\gamma| &= \sqrt{\frac{Cov(X, M)Cov(M, N)}{Cov(X, N)}} = \sqrt{\frac{\alpha\gamma\gamma\lambda}{\alpha\lambda}} \\ |\lambda| &= \sqrt{\frac{Cov(X, N)Cov(M, N)}{Cov(X, M)}} = \sqrt{\frac{\alpha\lambda\gamma\lambda}{\alpha\gamma}} \end{aligned}$$

for assessing identifiability, see [12]. Unfortunately Meek & Geiger's (1999) algorithm is not practical for even moderately-sized graphs.

Then we have:

$$\begin{aligned}
Cov(X,Y)Cov(X,M) &= \alpha^2\beta\gamma + \gamma\alpha\delta\theta \\
Cov(X,Y)Cov(X,M) - Cov(Y,M) &= \alpha^2\beta\gamma - \beta\gamma \\
\frac{Cov(X,Y)Cov(X,M) - Cov(Y,M)}{\sqrt{\frac{Cov(X,M)Cov(M,N)}{Cov(X,N)}}} &= \pm\ \alpha^2\beta - \beta \\
&= \pm\ \beta(\alpha^2 - 1) \\
\frac{Cov(X,Y)Cov(X,M) - Cov(Y,M)}{\sqrt{\frac{Cov(X,M)Cov(M,N)}{Cov(X,N)}}\left(\frac{Cov(X,M)Cov(X,N)}{Cov(M,N)} - 1\right)} &= \pm\ \beta
\end{aligned}$$

Thus, using only the observed covariances, we can identify both α and β up to the sign. Luckily, in medical contexts we have the intention-to-treat effect as well, which should tell us the sign of $\alpha\beta$.

The expression above is a very complicated function of several small numbers, each of which has its own estimation error, so it seems likely to be a very poor estimator. Note that we have two estimators for β – if we switch M with N in the expression above, we get a second estimator, so we could compare the estimates or take the average of the two. However, these estimators are not independent, and both are likely to have high variance. A good direction for future research would be to conduct simulation studies to explore the variability of these estimators (if analytic results are unavailable). My goal has been simply to show that unbiased estimation is possible; that the effect of interest can be identified, despite measurement error, if we assume a linear model.

12.3. **We cannot assume linearity for per protocol analyses.** Our motivation for exploring linear models was originally that measurement error destroys the d-separation relationships we rely on. Per protocol analyses also rely on d-separation relationships; they rely on *Adherence* screening off the back-door path through **U**. So the natural next step is to ask: can linear models let us estimate the effect of $Treatment$ on *Outcome* if the error-prone variable is *Adherence*, and we are attempting a per protocol analysis?

Unfortunately this move is forbidden by the deterministic relationships in the graph. The observed value of $Treatment$ is defined as the minimum of its two parents *Allocation* and the observed value of *Adherence* (assuming *Adherence* ranges from zero to one), so if there is measurement error in *Adherence* there will also be measurement error in $Treatment$. Thus, the graph will look different from normal representations of measurement error. $Treatment_{Measured}$ cannot be a child of the unobserved true value $Treatment$, because it is completely determined by *Allocation* and $Adherence_{Measured}$. Say Figure 21 represents the graph we'd get with perfect measurement – so the trial is perfectly blinded, but there is some non-adherence, and a back-door path from *Adherence* to *Outcomes* through unmeasured **U**. Then Figure 22 represents the situation with measurement error of *Adherence*. Because *Allocation* and $Adherence_M$ interact to determine the value of $Treatment_M$, we cannot assume a linear model. Some parts of if may be linear, but the $Adherence_M \rightarrow Treatment_M$ and $Allocation \rightarrow Treatment_M$ edges are not, and those

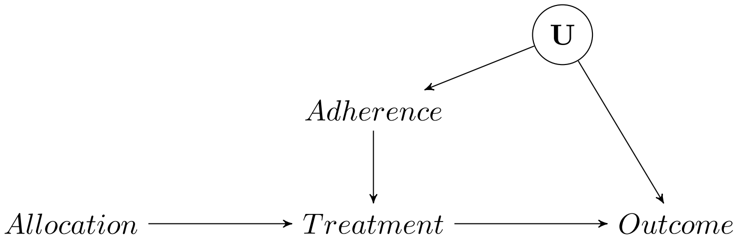

FIGURE 21. Without measurement error

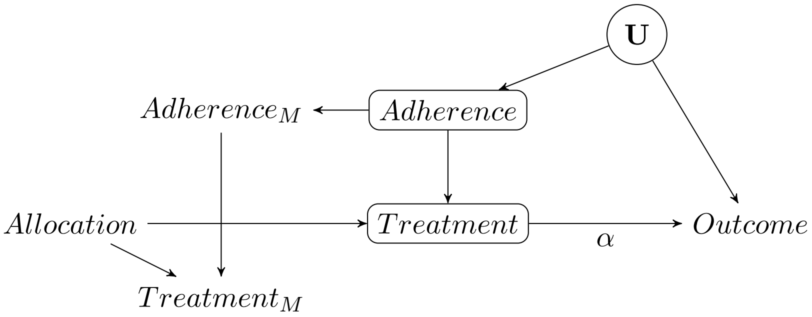

FIGURE 22. With measurement error of *Adherence*

edges are unavoidable if we are trying to calculate α (the *Treatment* → *Outcome* edge coefficient). Thus, we cannot take advantage of the Trek Rule to do a Per Protocol analysis if there is measurement error in *Adherence*.

Measurement error of *Adherence* is unbelievably common. Pill counts are often the only measure of adherence used in a trial, even though it has been shown that they overestimate adherence; furthermore, they do not capture all the causally relevant information in adherence behaviour, because they do not measure dose timing. [31] Likewise, measures of serum levels of drug overestimate adherence because of infrequent measurement and white-coat effects, and give very limited information about dose timing. [3] Self-report measures are only useful for those few patients who report non-adherence.

Electronic monitors are leagues ahead of all the other methods, although they have a slight tendency to underestimate adherence, thanks to some pocket-dosing behaviour. Data from electronic monitors has been used to model the serum concentration of drugs, indicating that they might be capable of capturing the causally relevant features of adherence. [39] Combined measures of adherence, using electronic monitors as well as some of the other methods, perform the best at predicting medical outcomes. [31] It is possible to measure adherence well; but it is rarely done.

Unless we have good measurement of adherence, a per protocol analysis will be biased by the confounding pathway through **U**. A good direction for future research would be to attempt to quantify the degree of the bias, and see whether data from electronic monitors, or perhaps a combined measure of adherence, can come close to eliminating the bias. If they can, electronic monitors should be used in every trial that will be analyzed by per protocol.

12.4. **Instrumental Variables.** It's trivial to show that we can identify the Treatment Effect using Instrumental Variables despite measurement error, provided the model is linear , and we have two different measures of *Treatment*.

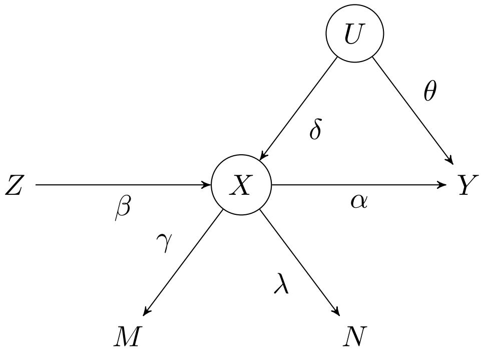

FIGURE 23. Instrumental Variables with measurement error, assuming a linear model

Figure 23 represents the causal structure, which give us the following constraints (among others):

$$Cov(Z, M) = \beta\gamma$$
$$Cov(Z, N) = \beta\lambda$$
$$Cov(Z, Y) = \beta\alpha$$
$$Cov(M, N) = \gamma\lambda$$

Thus we have:

$$\frac{Cov(Z, Y)\sqrt{Cov(M, N)}}{\sqrt{Cov(Z, M)Cov(Z, N)}} = \frac{\beta\alpha\sqrt{\gamma\lambda}}{\sqrt{\beta^2\gamma\lambda}} = \pm\, \alpha$$

We don't even need to use $Cov(Y, M)$ or $Cov(Y, N)$. Again, although we cannot identify the sign of α, the intention-to-treat estimate should do that for us. However, it is crucial that we have two measurements of X (which represents *Treatment* here). If we had measure M but not N, we would not be able to identify α.

Part 5. **Conclusion**

13. Summary and recommendations

The FDA and the Cochrane Collaboration prefer ITT over per protocol analyses. In other words, they care more about excluding *Adherence*-induced bias than they care about learning the Treatment Effect (in addition to the Allocation Effect). This demonstrates a degree of skepticism and caution towards adherence bias that is far higher than that directed toward bias induced by *Dropout*, and differential treatment and reporting biases. The FDA and the Cochrane Collaboration's preference holds even if there is no quantitative evidence, in a given trial, that bias induced by conditioning on *Adherence* will be larger than these other sources of bias. What makes *Adherence* bias different is that we can avoid dealing with it, because we can avoid conditioning on *Adherence*, whereas we are forced to condition on *Dropout* and we cannot fully prevent unblinding.

In any trial with substantial non-adherence, the Allocation Effect will be driven towards the null, even if the Treatment Effect is strong. In the case of the Women's Health Initiative trial of calcium and vitamin D, the ITT result emphasized in the abstract was non-significant, even though the subgroup analyses showed a statistically and clinically significant effect of treatment. The standard preference for ITT results led to overemphasis on the Allocation Effect – despite the fact that the Allocation Effect in the WHI may be very far from the allocation effect in clinical practice, given the high levels of vitamin D and calcium intake within the trial sample. The study authors considered the Treatment Effect, as measured by per protocol analysis, to be more relevant to clinical practice. If we always let concerns about *Adherence* bias trump concerns about relevance, regardless of whether the risk of bias were high or low, we would recommend that women not bother taking calcium and vitamin D. This would be a mistake.

I have argued that the FDA and the Cochrane Collaboration's preference stems partly from the misinterpretation of a historic trial, the Coronary Drug Project (CDP). This trial is cited in the Cochrane Collaboration's handbook in support of their preference for ITT over per protocol analyses; it was also cited by Russell Katz, spokesperson for the FDA, when he explained the FDA's preference for ITT. The CDP demonstrated a strong association between *Adherence* and *Outcome*, but did not demonstrate that *Allocation* had any effect on *Adherence*; in fact, evidence from the trial seems to suggest the opposite. The CDP research group argued that this edge was possible in principle (because *Allocation* precedes *Adherence* in time). They looked at the strong association between *Adherence* and *Outcomes* produced by the $\textit{Adherence} \leftarrow \mathbf{U} \rightarrow \textit{Outcomes}$ pathway, and inferred that there might be large bias introduced through $\textit{Allocation} \rightarrow \textit{Adherence} \leftarrow \mathbf{U} \rightarrow \textit{Outcomes}$ if they conditioned on *Adherence*. I believe the data from the CDP demonstrate that there was no such bias, and that it would be more productive to check for the presence of bias rather than ruling out per protocol analyses *a priori*. I also believe that the lack of graphical representations of the problem allowed the two pathways to be conflated; the graphical representations I develop in Part 2 should allow for clearer thinking about the sources of bias in clinical trials.

Given the graphical representations of clinical trials, it is clear that we can take several different approaches to measuring and mitigating bias when we estimate the treatment effect. Firstly, we can try to block the pathways that cause bias by measuring variables that are likely to be on those pathways. There are several candidates for members of **U**. We could also measure a $Mediator$ variable between $Treatment$ and *Outcome*, and use it to perform a front door analysis, in which case we would not need to assume blinding at all.

Secondly, we can assess the degree of unblinding by measuring $Beliefs$ directly. If for some reason we cannot measure $Beliefs$, we can use relationships among the measured variables to assess the likelihood of unblinding. In a blinded trial, we would expect *Allocation* to be independent of $Beliefs$, *Adherence*, *Dropout*, *Adverse effects*, and any $Outcome$ that is apparent to the participant. In cases like the CDP's trial of clofibrate, where all evidence points to the blind holding, we should allow per protocol analyses.

If *Allocation* does appear to affect *Adherence* but the assumptions of an Instrumental Variables (IV) analysis hold, then we can use instrumental variables to estimate the treatment effect. IV assumptions require that there be no direct effect of *Allocation* on $Outcome$, and they require one of several possible parametric assumptions: either the model should be linear, or log-linear, or monotonic, or else the d-separation relationships among the variables should hold when $Treatment$ and *Outcome* are made binary.

Lastly, I show that when measurement error is suspected, we can still identify the effect of treatment using the front door method, so long as we can assume linearity, and we have made two measurements of the $Mediator$ variable. If there is measurement error in the measurement of *Adherence*, an instrumental variables analysis can identify the effect using two measures of *Adherence* and assuming linearity. However, the assumption of linearity is inappropriate for per protocol analyses, because of the non-linear relationship between $Treatment$ and its parents, *Allocation* and *Adherence*. For this reason, if researchers wish to perform a per protocol analysis, it is crucial that they measure adherence as well as possible, which means using electronic monitors.

In sum, there is no reason to give up on measuring the Treatment Effect. It need not always take second place to the Allocation Effect, as measured by an ITT analysis. There are many ways to measure and mitigate the bias introduced in analyses of the treatment effect; some decisions can be made at the analysis stage, but several steps must be taken when the trial is designed, such as using electronic monitors rather than pill counts to measure adherence. The one step I recommend most highly it is to measure participants' and assessors' $Beliefs$ about allocation. Unblinding is the source of all our analytic woes, so we should learn whether it occurred or not.

14. Further research

14.1. **Simulation studies.** It would be extremely helpful to perform simulation studies of all these different analytic techniques, to see how biased they are under different plausible assumptions about the strength of the confounding pathways. It would also be helpful to see simulations of the estimators I propose for in the measurement error section. Although they are unbiased, they are likely to have very high variance, and it would be worthwhile to check whether the estimates they produce are worse than the biased estimates produced by the unadjusted method.

14.2. **Vested interests.** One of the implicit assumptions I make throughout this thesis is that every epistemic agent is acting in good faith. If the treatment is ineffective (or dangerous), we all want to learn that it is ineffective (or dangerous). In the field of pharmaceutical trials this assumption is laughably false. Many if not most of the researchers have vested interests in the results of their experiments. The pharmaceutical industry's power over the FDA [1, pages 208–14] makes thorough regulation impossible, but the agency still battles to prevent regulatory mistakes. In this climate inference is not a game against nature or stochasticity; it is instead an adversarial process, and in that sense is more akin to legal than scientific inference.

ITT analyses have been praised for being "conservative" [30][24] – i.e. they make it harder to demonstrate treatment efficacy – on the grounds that approving an ineffective drug has worse consequences than missing an effective one. But the relative utility of the two options is never properly calculated and compared. I believe the preference for "conservative" (read: biased in a particular direction) analyses stems from the adversarial climate in medical research. As a result, purely statistical considerations do not capture the constraints on inference, and this area would benefit from the expertise of legal epistemologists.

References

[1] Marcia Angell. *The truth about drug companies: how they deceive us and what to do about it.* Scribe Publications, Carlton North, Victoria, Australia, revised (2006) edition, 2005.

[2] Stuart G Baker and Barnett S Kramer. Randomized trials for the real world: making as few and as reasonable assumptions as possible. *Statistical Methods in Medical Research*, 17:243–252, 2007.

[3] Karina M. Berg and Julia H. Arnsten. Practical and conceptual challenges in measuring antiretroviral adherence. *J Acquir Immune Defic Syndr*, 43(Suppl. 1):S79–S87, December 2006.

[4] Isabelle Boutron, Candice Estellat, and Philippe Ravaud. A review of blinding in randomized controlled trials found results inconsistent and questionable. *Journal of Clinical Epidemiology*, 58:1220–1226, 2005.

[5] Adam La Caze. *Evidence Based Medicine: Evolution, Revolution, or Illusion?: A philosophical examination of the foundations of Evidence Based Medicine.* Doctor of philosophy, Department of Philosophy, The University of Sydney, 2008.

[6] Richard H Chapman, Joshua S Benner, Allison A Petrilla, Jonothan C Tierce, S Robert Collins, David S Battleman, and J Sanford Schwatz. Predictors of adherence with antihypertensive and lipid-lowering therapy. *Archives of Internal Medicine*, 165:1147–1152, 2005.

[7] Rory Collins and Stephen MacMahon. Reliable assessment of the effects of treatment on mortality and major morbidity, i: clinical trials. *Lancet*, 357:373–380, 2001.

[8] P J Devereaux, Mohit Bhandari, Victor M Montori, Braden J Manns, William A Ghali, and Gordon H Guyatt. "double blind, you are the weakest link – goodbye!". *Evidence Based Nursing*, 5:36–37, 2002.

[9] B. Efron and D. Feldman. Compliance as an explanatory variable in clinical trials. *Journal of the American Statistical Association*, 86(413):9–26, March 1991.

[10] Dean Fergusson, Kathleen Cranley Glass, Duff Waring, and Stan Shapiro. Turning a blind eye: the success of blinding reported in a random sample of randomised, placebo controlled trials. *British Medical Journal*, 328:432 (doi:10.1136/bmj.37952.631667.EE), January 2004.

[11] Food and Drug Administration. *Guidance for Industry: E 10. Choice of Control Group and Related Issues in Clinical Trials*, 2001. Available from: `http://www.fda.gov/downloads/RegulatoryInformation/Guidances/ucm125912.pdf` [cited June 6, 2013].

[24] At least in placebo-controlled (superiority) trials, when evaluating efficacy. In active-controlled (equivalence) trials, or when evaluating adverse effects, ITT is anti-conservative.

[12] Dan Geiger and Chris Meek. Quantifier elimination for statistical problems. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 226–235. Morgan Kaufmann Publishers Inc., July 1999.

[13] M Maria Glymour and Sander Greenland. Causal diagrams. In Kenneth J Rothman, Sander Greenland, and Timothy L Lash, editors, *Modern Epidemiology*, chapter 12, pages 183–209. Lippincott Williams and Wilkins, 530 Walnut St, Philadelphia, PA 19106 USA, 3rd edition, 2008.

[14] Sander Greenland. Quantifying biases in causal models: Classical confounding vs collider-stratification bias. *Epidemiology*, 14(3):300–306, 2003.

[15] Jason Grossman and Fiona J. Mackenzie. The randomized controlled trial: gold standard, or merely standard? *Perspectives in Biology and Medicine*, 48(4):516–534, 2005.

[16] Anturane Reinfaction Trial Research Group. Sulfinpyrazone in the prevention of cardiac death after myocardial infarction. *New England Journal of Medicine*, 298(6):289–295, 1978.

[17] Anturane Reinfaction Trial Research Group. Sulfinpyrazone in the prevention of sudden death after myocardial infarction. *New England Journal of Medicine*, 302(5):250–256, 1980.

[18] Coronary Drug Project Research Group. The coronary drug project: design, methods and baseline results. *Circulation*, 47 & 48(Supplement 1):I–1–I–179, 1973.

[19] Coronary Drug Project Research Group. Clofibrate and niacin in coronary heart disease. *Journal of the American Medical Association*, 231:360–381, 1975.

[20] Coronary Drug Project Research Group. Influence of adherence to treatment and response of cholesterol on mortality in the coronary drug project. *New England Journal of Medicine*, 303(18):1038–1041, October 1980.

[21] Miguel A. Hernán and James M. Robins. Instruments for causal inference: An epidemiologist's dream? *Epidemiology*, 17:360–372, 2006.

[22] J P T Higgins and S Green, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. The Cochrane Collaboration, version 5.1.0 edition, March 2011.

[23] Sally Hollis and Fiona Campbell. What is meant by intention to treat analysis? survey of published randomised controlled trials. *British Medical Journal*, 319:670–674, September 1999.

[24] Ralph I. Horwitz and Sarah M Horwitz. Adherence to treatment and health outcomes. *Archives of Internal Medicine*, 153:1863–1868, August 1993.

[25] Rebecca D. Jackson, Andrea Z. LaCroix, Margery Gass, Robert B. Wallace, John Robbins, Cora E. Lewis, Tamsen Bassford, Shirley A.A. Beresford, Henry R. Black, Patricia Blanchette, Denise E. Bonds, Robert L. Brunner, Robert G. Brzyski, Bette Caan, Jane A. Cauley, Rowan T. Chlebowski, Steven R. Cummings, Iris Granek, Jennifer Hays, Gerardo Heiss, Susan L. Hendrix, Barbara V. Howard, Judith Hsia, F. Allan Hubbell, Karen C. Johnson, Howard Judd, Jane Morley Kotchen, Lewis H. Kuller, Robert D. Langer, Norman L. Lasser, Marian C. Limacher, Shari Ludlam, JoAnn E. Manson, Karen L. Margolis, Joan McGowan, Judith K. Ockene, Mary Jo O'Sullivan, Lawrence Phillips, Ross L. Prentice, Gloria E. Sarto, Marcia L. Stefanick, Linda Van Horn, Jean Wactawski-Wende, Evelyn Whitlock, Garnet L. Anderson, Annlouise R. Assaf, and David Barad. Calcium plus vitamin d supplementation and the risk of fractures. *New England Journal of Medicine*, 354(7):669–683, February 2006.

[26] Leslie Kamen-Siegel, Judith Rodin, Martin E. P. Seligman, and John Dwyer. Explanatory style and cell-mediated immunity in elderly men and women. *Helth Psychology*, 10(4):229–235, 1991.

[27] Russell Katz. Regulatory view: Use of subgroup data for determination of efficacy. In J A Cramer and B Spilker, editors, *Patient compliance in medical practice and clinical trials*, chapter 21, pages 251–263. Raven Press, Ltd., New York, 1991.

[28] Gunver S. Keinle and Helmut Keine. The powerful placebo effect: Fact or fiction? *Journal of Clinical Epidemiology*, 50(12):1311–1318, 1997.

[29] J Kolahi, H Bang, and J Park. Towards a proposal for assessment of blinding success in clinical trials: up-to-date review. *Community Dent Oral Epidemiol*, 37:477–484, 2009.

[30] J A Lewis and D Machin. Intention to treat – who should use itt? *British Journal of Cancer*, 68:647–650, 1993.

[31] Honghu Liu, Carol E. Golin, Loren G. Miller, Ron D. Hays, C. Keith Beck, Sam Sanandaji, Judith Christian, Tomasa Maldonado, Dena Duran, Andrew H. Kaplan, and Neil S. Wenger. A comparison study of multiple measures of adherence to hiv protease inhibitors. *Annals of Internal Medicine*, 134:968–977, 2001.

[32] Honghu Liu, Loren G. Miller, Ron D. Hays, Carol E. Golin, Tongtong Wu, Neil S. Wenger, and Andrew H. Kaplan. Repeated measures longitudinal analyses of hiv virologic response as a function of percent adherence, dose timing, genotypic sensitivity, and other factors. *Journal of Acquired Immune Deficiency Syndrome*, 41:315–322, 2006.

[33] Graham S May, David L DeMets, Lawrence M Friedman, Curt Furberg, and Eugene Passamani. The randomized clinical trial: bias in analysis. *Circulation*, 64(4):669–673, 1981.

[34] Victor M. Montori and Gordon H. Guyatt. Intention-to-treat principle. *Canadian Medical Association Journal*, 165(10):1339–1341, 2001.

[35] J H Noseworthy, G C Ebers, M K Vandervoort, R E Farquhar, E Yetisir, and R Roberts. The impact of blinding on the results of a randomized, placebo-controlled multiple sclerosis clinical trial. *Neurology*, 44:16–20, 1994.

[36] Judea Pearl. Causal inference in the health sciences: A conceptual introduction. *Health Services & Outcomes Research Methodology*, 2:189–220, 2001.

[37] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2nd edition, September (forthcoming) 2009.

[38] Peter M Rothwell. External validity of randomised controlled trials: “to whom do the results of this trial apply?”. *Lancet*, 365:82–93, 2005.

[39] Ana Rubio, Christopher Cox, and Michael Weintraub. Prediction of diltiazem plasma concentration curves from limited measurements using compliance data. *Clinical Pharmacokinetics*, 22(3):238–246, 1992.

[40] Kenneth F Schulz, Iain Chalmers, Richard J Hayes, and Douglas G Altman. Empirical evidence of bias: Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. *Journal of the American Medical Association*, 273(5):408–412, February 1 1995.

[41] Kenneth F Schulz and David A Grimes. Blinding in randomised trials: hiding who got what. *Lancet*, 359:696–700, 2002.

[42] Daniel Schwartz and Joseph Lellouch. Explanatory and pragmatic attitudes in therapeutic trials. *Journal of Chronic Diseases*, 20:637–648, 1967.

[43] Cosma Shalizi. Advanced data analysis from an elementary point of view [online]. Available from: `http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/ADAfaEPoV.pdf`.

[44] Sol Sherry. The anturane reinfarction trial. *Circulation*, 62(Supp V):V73–V78, December 1980.

[45] Scot H Simpson, Dean T Eurich, Sumit R Majumdar, Rajdeep S Padwal, Ross T Tsuyuki, Janice Varney, and Jeffrey A Johnson. A meta-analysis of the association between adherence to drug therapy and mortality. *British Medical Journal*, 333:15–20, June 2006.

[46] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction and Search*. MIT Press, Cambridge, Massachusetts, 2nd ed. edition, 2000.

[47] S. Sullivant, K. Talaska, and J. Draisma. Trek separation for gaussian graphical models. *Annals of Statistics*, 38(3):1665–1685, 2010.

[48] Robert Temple and Gordon W Pledger. Special report: The fda’s critique of the anturane reinfaction trial. *New England Journal of Medicine*, 303(25):1488–1492, December 1980.

[49] John Urquhart. Patient compliance as an explanatory variable in four selected cardiovascular studies. In J A Cramer and B Spilker, editors, *Patient compliance in medical practice and clinical trials*, chapter 24, pages 301–322. Raven Press, Ltd., New York, 1991.

[50] John Urquhart. Variable patient compliance in ambulatory trials: nuisance, threat, opportunity. *Journal of Antimicrobial Chemotherapy*, 32:643–649, 1993.

[51] John Urquhart. Pharmionics: research on what patients do with prescription drugs. *Pharmacoepidemiology and Drug Safety*, 13:587–590, 2004.

[52] Bernard Vrijens and John Urquhart. Patient adherence to prescribed antimicrobial drug dosing regimens. *Journal of Antimicrobial Chemotherapy*, 55:616–627, 2005.