make predictions. However, if you do not know the height of a plant or the sex of a person or their blood pressure when deciding on a treatment, then you cannot use them. In such cases, the appropriate predictive procedure depends crucially on how the data were obtained. If predictor information becomes available later, and you can change treatments, you might want to do that.

While prediction is the ultimate goal of science, causation is the warm fuzzy. Causation can greatly simplify prediction and we like to think that good causative models provide the best predictions. But in the end, getting predictions correct is more important than imagining that we understand why things happen the way they do. While I admit that I am not an expert on the causal model literature, I am unfamiliar with any satisfactory way to infer causation other than performing randomized experiments. Sure, data analysis can help you choose between two or more causative models, but that is a far cry from inferring causation from data analysis. In fact, without knowing the sampling design, we cannot even be sure of making appropriate predictions from data analysis alone.

REFERENCES

Aldrich, J. (2005), "Fisher and Regression," *Statistical Science*, 20, 401–417. [14]

Baker, S. G. (2013), "Causal Inference, Probability Theory, and Graphical Insights," *Statistics in Medicine*, 32, 4319–4330. [15]

Christensen, R. (1997), *Log-Linear Models and Logistic Regression* (2nd ed.), New York: Springer. [13,14]

Cox, D. R. (1958), *Planning of Experiments*, New York: Wiley. [15]

Spirtes, P., Glymour, C., and Scheines, R. (2000), *Causation, Prediction, and Search* (2nd ed.), Cambridge: MIT Press. [13]

Comment: A Fruitful Resolution to Simpson's Paradox via Multiresolution Inference

Keli LIU and Xiao-Li MENG

Simpson's Paradox is really a Simple Paradox if one at all. Peeling away the paradox is as easy (or hard) as avoiding a comparison of apples and oranges, a concept requiring no mention of causality. We show how the commonly adopted notation has committed the gross-ery mistake of tagging unlike fruit with alike labels. Hence, the "fruitful" question to ask is not "Do we condition on the third variable?" but rather "Are two fruits, which appear similar, actually similar at their core?." We introduce the concept of *intrinsic* similarity to escape this bind. The notion of "core" depends on how deep one looks—the multi resolution inference framework provides a natural way to define intrinsic similarity at the resolution appropriate for the treatment. To harvest the fruits of this insight, we will need to estimate intrinsic similarity, which often results in an indirect conditioning on the "third variable." A ripening estimation theory shows that the standard treatment comparisons, unconditional or conditional on the third variable, are low hanging fruit but often rotten. We pose assumptions to pluck away higher-resolution (more conditional) comparisons—the multiresolution framework allows us to rigorously assess the price of these assumptions against the resulting yield. One such assessment gives us *Simpson's Warning: less conditioning is most likely to lead to serious bias when Simpson's Paradox appears*.

KEY WORDS: Bias-variance tradeoff; Principal stratification

1. THE SOURCE OF CONFUSIONS AND DEBATES

1.1 Comparing Apples and Oranges

Imagine Ms. Broken going to Dr. Heal to be treated for heart disease. A new treatment was made available to Dr. Heal, who also learned from a clinical trial that it can substantially outperform a standard treatment used as its control. However, its effectiveness depends on a patient's cholesterol level, which can also be altered significantly by the treatment. Therefore, to determine the appropriate treatment for Ms. Broken, Dr. Heal needs to know how trial subjects with cholesterol level similar to Ms. Broken's (say about 240 mg/dL) responded to the

Keli Liu (E-mail: *keli.liu25@gmail.com*) is A.B. Graduate, and Xiao-Li Meng (E-mail: *meng@stat.harvard.edu*) is Whipple V. N. Jones Professor, Department of Statistics, Harvard University, Cambridge, MA 02138. The authors thank Dr. Armistead for an invigorating article that stimulated authors to take a higher resolution look at Simpson's Paradox. Jessica Hwang provided invaluable advice on organization and structure, as well as pointing out incoherencies and inconsistencies. Of course, the remaining incoherencies and inconsistencies are entirely of the authors. The authors also thank the NSF for partial financial support, and *TAS* editor, Ronald Christensen, and journal manager, Eric Sampson, for their saintly patience amid growing despair. During the preparation of this discussion, we learned the sad news of Dennis Lindley's passing. Without his work, our understanding of this and many other topics would not be as rich today. We therefore dedicate this article in his memory.

two treatments. The clinical trial data did include cholesterol measurements for patients. Dr. Heal therefore seems to face a rather simple task: compare the treatment effects and side effects among those in the clinical trial with cholesterol level at 240 mg/dL. Or using the terminology of Dr. Armistead's stimulating article, Dr. Heal needs to condition on the *third variable*, cholesterol level, $Z = 240$, when comparing the outcome variable Y (e.g., $Y = 1$ indicates success and $Y = 0$ otherwise) between the treatment group (indicated by $T = 1$) and the control group ($T = 0$).

So why all the fuss about whether or not to condition on the third variable Z? Since conditioning always leads to a more refined state space, should not we always condition, at least in theory? The answer is yes, provided that we condition on the *right conditions*. In the scenario above, we were deliberately vague about "cholesterol measurements of patients," with the measurement time unspecified. Clearly to be relevant for Ms. Broken's choice, Dr. Heal should condition on cholesterol measurements taken at or just prior to the treatment. But what if the available data are *post*-treatment measurements? Suppose the new treatment decreases cholesterol by 20 mg/dL, whereas the standard treatment has little impact. Then $Z_{\text{post}} = 240$ corresponds to $Z_{\text{at}} = 260$ for the $T = 1$ group, clearly incomparable with $Z_{\text{at}} = Z_{\text{post}} = 240$ for the $T = 0$ group (assuming the treatment period is short enough that temporal effects are negligible). Therefore, conditioning on the same value of Z_{post} actually leads to a comparison of apples and oranges: individuals alike *post*-treatment may be highly unlike *at*-treatment. This illustrates the rationale for Pearl's emphasis to not condition on variables affected by treatment.

However, the logical implication of "do not condition on variables affected by treatment" is *not* "condition on nothing." In our example, clearly we should condition on Z_{at}, which, though unmeasured, could be deduced from Z_{post} if we knew how the treatments acted on cholesterol. Thus, even if we should not condition on Z_{post} itself, we should condition on a *function* of it. *Third variables* affected by treatments are therefore not useless; it is just that an extra processing step is required.

In practice, we typically do not have full knowledge about how Z_{at} impacts Z_{post}. However, as we shall demonstrate, even weak information or assumptions can lead to substantively higher quality comparisons than not conditioning (or improper conditioning). Our emphasis therefore is not to decide whether or when we should condition or not. Rather, we focus on the following more productive questions, and explore how the framework of multiresolution inference (Meng 2014) can help to answer them.

I. *Ideal Question*. If we had all the data we wished, what is the *ideal* (infinite resolution) conditioning that allows us to answer the substantive question exactly?
II. *Inferential Question*. How can we best approximate the ideal conditioning by an *operational* (finite resolution) conditioning (which may still not be directly estimable from the data)?
III. *Estimation Question*. How can we best estimate our operational conditioning based on the data we observe (data resolution)?

1.2 In Case Causality is Not Your Cup of Fruit...

Before we proceed, we echo Dr. Armistead's sentiment that the central issues of Simpson's Paradox can be addressed adequately without necessarily invoking causality, unless one takes an encompassing view that no inference is complete without stating its causal origins or consequences. Causality is a useful *tool* for answering question (I), helping to define and identify the ideal "at-treatment" characteristics Z_{at} (Pearl 2000). But it is not the only tool (see Section 2), and its overhead (e.g., familiarity with causal diagrams) may mask for some the fundamental motivation behind using it in the first place: avoid comparing apples and oranges. Besides, it provides little help for answering (II) or (III), which asks for a meaningful way to reduce the ideal set of Z_{at} to meet practical constraints. The easy answers, "include all Z_{at}" or "include all estimable Z_{at}," turn out to be inadequate, as we shall demonstrate.

A moment's reflection on questions (I)–(III) shows that the inferential question is really a question about striking a balance between what we want to know (I) and what we can answer (III). Suppose you are standing at the edge of a lava pit with a treasure on an island in the pit's center. Pillars of obsidian jut out from the lava, and you can only jump on one before vaulting onto the island. But which one? Too close to the center may lead to a fiery death on your first jump; too far, and the same fate awaits on your second jump. This is precisely the "deathtrap" we face in choosing an operational conditioning. Yet mention of this tradeoff is absent from the usual discussions on Simpson's Paradox. Why is this?

The controversy surrounding "the third variable" has raged because we often focus on one of two objectives to the exclusion of the other. The first objective seeks to identify the types of variables we can *include* in the operational conditioning. The second to identify the operational conditions we can *estimate* from the data. Those who seek only the first objective would conclude that causality solves everything. Others who worry only about the second would wonder why causality matters at all. But there is a third objective—decide what variables we *should* include in the operational conditioning—which is the most important and the logical capstone to the first two. Therefore, in the spirit of Dr. Armistead's article, let us work toward a rebirth. But instead of the third variable, let us resurrect the "third objective!" The multiresolution setup in separating the ideal, the inferential, and the estimation questions directly addresses this forgotten objective of "should."

2. A RESOLUTION VIA MULTIPLE RESOLUTION

Our motivating example shows that we can and should use Z to infer the *at-treatment similarity* between subjects. While intuitive, the notion of at-treatment is clunky in that it suggests our reasoning depends on a temporal structure. In addition, all individuals are fundamentally unique, so what does *similar* mean? A formalism comprising the potential outcomes framework (see Rubin 2005) and the multiresolution framework (see Meng 2014) turns out to be adequate for formulating the meaning of at-treatment similarity. Wasserman (2013, June 20, Blog) argued that the potential outcomes framework (or some causal

variant) is necessary for understanding Simpson's Paradox. The gist of the matter is that we need to distinguish between the logical statements we *wish* to make (which are in terms of potential outcomes), and the probabilistic statements that we can *estimate* (which are in terms of observed data). This parallels our caveat above to distinguish the estimand Z_{at} from estimators based on Z. However, we would be complacent to simply warn against conditioning on Z directly. We need to know how to condition on Z *indirectly* but correctly. The multiresolution framework addresses this deficiency.

2.1 The Ideal Question: Infinite Resolution for Individuality

Let Ω be our population of interest. Each $\omega \in \Omega$ represents an individual. All the *intrinsic characteristics* of this individual (e.g., age, sex, genomic signature, etc., when ω represents a human) are encoded into ω. The idea behind the potential outcome framework is to imagine copies of these individuals in parallel universes where they receive different treatments. Similar to the setup in Pearl (2011), this can be formalized mathematically by considering an augmented *product* space $\Omega^A \equiv \Omega \times \mathbb{T}$, where $\mathbb{T}$ is the space of treatment assignments. A common setting is $\mathbb{T} = \{0, 1\}$, as in our motivating example. We can never study any individual ω *in isolation*: our data are always produced from the individual *in some universe*. As a helpful analogy, one might think of ω as the Platonic form of the *realized state* (ω, t) (of course, we cannot access the Platonic form directly). Anything we can observe are functions of the realized state, that is, $f(\omega, t)$, which could happen to be free of t but such are special cases. The distinction between ω and (ω, t) is fundamental to what follows—as our invocation of Plato might suggest, what we really care about are properties of ω, the form, and not properties of (ω, t), the realization of the form in a particular universe. To make this distinction clear, we will refer to ω as the individual and to (ω, t) as the state or realized state.

It is essential to understand that the assumption of the product space implies a decoupling between individuality, ω, and the treatment. Hence, intrinsic characteristics encoded in ω must remain invariant to treatment. The white–black plants example in Dr. Armistead's article is useful for illustrating this point. Is color a treatment *for the individual plant*—is Ω in this case the population of plants? The product space assumption, $\Omega \times \mathbb{C}$ (where $\mathbb{C}$ is {white, black}), says that we can choose any plant, ω, and the state (ω, c) *must be realizable in the world for all* c. But color, $C(\omega)$, is purely a function of ω, so the only realizable state is $(\omega, C(\omega))$. Hence, the product space assumption, $\Omega \times \mathbb{C}$, is violated if Ω is the population of plants. Then what is Ω in this case? Suppose that there is a 1–1 correspondence between the color of a plant and the first base pair in the plants genome, g (a vector of base pairs). Suppose further that the support of possible g is the product of the supports of g's components (this does not hold in practice but is used here for simplicity of illustration). Then we can decompose g as $g = (c, g_{-1})$, where g_{-1} is g with the first component removed. We can now let $\omega = g_{-1}$, so that Ω is the population of "proto"-plants. If we conceive of treatment as color, our unit of analysis is no longer plants but proto-plants. *Any* variable can be conceived of as a treatment, as long as we correctly identify the "individual" (and population) for which that variable is a treatment.

Having determined what our "individual" is, we can now ask whether we should condition on plant height when our treatment is color. As we discuss below, we should only condition on characteristics *intrinsic* to our individual—the appropriate conditioning depends on the unit of analysis. In this case, individual means proto-plant *not* plant. In the data, color is associated with plant height. *Assuming* that nature generated our sample of plants through independent sampling of proto-plants and colors, that is, assuming a randomized experiment with proto-plants as units, our discussion below shows that this association implies that plant height functionally depends on color. Proto-plants lack color, hence, plant height cannot be an intrinsic characteristic of proto-plants. This is not to say that there is not another characteristic called "proto-plant height" that is an intrinsic characteristic (and which we should condition on), but proto-plant height is different from the measured height (which is plant height). Conditioning on the latter does not lead to conditioning on the former.

In general, we do not know how nature creates plants from proto-plants and colors—that is to say when we conceive of the treatment as color, we lack information on whether or not the observed data can be analyzed as a randomized experiment with proto-plants as the unit (or whether it should be seen as an observational study). Thus even though *conceptually*, comparing the effect of pesticides (with plants as the individual) is no different than comparing the effect of color (with proto-plants as the individual), having changed the unit of analysis, we lose (or gain) information on the generation of the data. Hence, comparing the effect of color is *practically* more challenging, because *nature* and not the scientist designs the experiment. This explains why we are fundamentally uncomfortable with seeing color as a "treatment" despite the fact that any variable is a treatment for some definition of "individual." It is not a treatment that *we* can apply.

To further emphasize the role played by the unit of analysis, note that each proto-plant, ω, defines an equivalence class in the population of plants (one containing all plants with identical genome except possibly in the first base pair). So our unit of analysis is an individual when the population consists of proto-plants and an equivalence class when the population comprises plants. Clearly, proto-plant is a *lower-resolution* unit of analysis than a plant. Hence, we can say that a treatment applied to individual plants (e.g., spraying pesticide) is of *higher resolution* than a treatment applied to equivalence classes of plants or proto-plants (e.g., color). The key to understanding what follows is that the correct conditioning should match the resolution of the treatment. In fact, another way to say "do not condition on variables affected by treatment" is "do not condition on characteristics that exceed the resolution of the treatment." That is, plant height, a characteristic of individual plants, exceeds the resolution of the treatment (color), which is applied to an equivalence class of plants. The multiresolution perspective presents this crucial insight in its most transparent form: *know your unit of analysis*.

Whereas our setup is generic, for concreteness of discussion, we will focus on human populations and let $Y(\omega, t) \in \{0, 1\}$ be the health status (e.g., cured or uncured) of state (ω, t). When an individual, ω^*, walks into a doctor's office, ideally the doctor makes a choice of treatment by comparing $Y(\omega^*, 0)$ with $Y(\omega^*, 1)$, which obviously are unavailable. However, this lack of direct data does not and should not deter us from formulating the question as the doctor asks it. Only after formulating the correct ideal question can we formulate the relevant inferential question. Throughout, we will also assume that the data are generated from a randomized experiment: the states (ω, t) that comprise our data are sampled through independent sampling of ω and t. Mathematically, this means that the treatment T will be functionally independent of any function of ω alone, $f(\omega)$, a critical assumption for the discussions below.

2.2 The Inferential Question: Finite Resolution for Similarity

Our fundamental inference challenge—as Lindley and Novick (1981) stated—is to make a valid statement about $(Y(\omega^*, 0), Y(\omega^*, 1))$ when we only observe $Y(\omega, 0)$ and $Y(\omega', 1)$ for some $\omega, \omega' \neq \omega^*$. To address this problem, Lindley and Novick (1981) relied on the concept of exchangeability. We prefer the notion of *resolution*. The change in terminology suggests a different emphasis *in action*. Exchangeability is something we assume; resolution is something we can adjust. When we have no *direct data* (exact replications) to learn $Y(\omega^*, 0)$ or $Y(\omega^*, 1)$, we say that the resolution of the estimand, $(Y(\omega^*, 0), Y(\omega^*, 1))$, exceeds the resolution of our dataset. In such cases, which include all clinical trials, we have to create approximate clones for ω^*. The observed states of these approximate clones then provide *indirect data* (approximate replications) with which to infer $Y(\omega^*, 0)$ and $Y(\omega^*, 1)$. The resolution of our inference can be thought of as how strict we are in letting some ω be an approximate clone of ω^*. The inferential question is: how strict should we be?

In finding an optimal strictness, we need to account for the error of approximating ω^* by ω. How much do ω and ω^* differ with respect to *intrinsic characteristics*—those depending only on ω and ω^*, and not their particular states, (ω, t) and (ω^*, t^*)? As any good scientist would do, we wish to compare states with different treatments but *intrinsically similar* individuals. We are simply formalizing the scientific idea of *ceteris paribus*, holding all else (ω) constant except the treatment assignment. For $C(\omega, t)$, a *realized-state characteristic*, to also be an intrinsic characteristic, we must have $C(\omega, 0) = C(\omega, 1)$. Thus, we say that a nonconstant (vector) function C, defined on $\Omega^A = \Omega \times \mathbb{T}$, records an intrinsic characteristic of ω if $C(\omega, t)$ can be written as $C(\omega)$, that is, C is *functionally independent* of the treatment.

Equipped with this definition, we say that ω is an approximate clone for ω^*, if for a selected set of intrinsic characteristics, $C(\omega) = C(\omega^*)$. That is, we define the *ω^*-relevant subpopulation with respect to C* as $\Omega_C(\omega^*) = \{\omega : C(\omega) = C(\omega^*)\}$. The resolution level, R, can then be defined as a numerical index of how restrictive this subpopulation is. A convenient choice is the dimension of C, with infinite resolution corresponding to cases where the only acceptable clone of ω^* is itself; see Meng (2014). Though flawed in many regards (e.g., it does not distinguish different kinds of infinite resolutions), this intuitive notion of resolution will suffice for our discussion of Simpson's Paradox. Our whole point is that it is wrong to define $\Omega_C(\omega^*)$ using a realized-state characteristic $C(\omega, t)$ that *functionally depends* on t—this mistakes the fundamental unit of analysis as (ω, t) rather than ω. As Plato would remind us, we should not care about superficial similarities $(C(w, t) = C(w', t'))$ but rather about intrinsic similarities $(C(w) = C(w'))$.

Example 1. Is it possible to define intrinsic similarity via the notion of independence most familiar to statisticians, that is, *stochastic independence*? Intuitively, if "T does not affect Z," then Z is a property intrinsic to the individual rather than a consequence of treatment. This intuition is indeed correct, but equating "does not affect" with *stochastic independence* is not. To see this, let Z_{at} be the *standardized* cholesterol level (in a population of interest) at the time of treatment such that $Z_{\text{at}} \sim N(0, 1)$. Suppose the standardized post-treatment cholesterol level, Z, is linked to Z_{at} via

$$Z = T(-Z_{\text{at}}) + (1 - T)(Z_{\text{at}}) = (1 - 2T)Z_{\text{at}}. \tag{1}$$

Because $Z|T \sim N(0, 1)$, Z is properly standardized *conditionally and unconditionally.* Consequently, Z is stochastically independent of T. Yet when we condition on $Z = z$ in the treatment group $T = 1$, we obtain the subpopulation where $Z_{\text{at}} = -z$. In the control group $T = 0$, however, restricting $Z = z$ would lead to the subpopulation where $Z_{\text{at}} = z$, clearly a rather different subpopulation from the one for $T = 1$.

So what does stochastic independence give us? The stochastic independence of T and Z guarantees that Simpson's Paradox does not occur (Wasserman 2013, June 20, Blog). The signs of the two comparisons, conditioning on Z or not, will then agree. But this agreement itself says little about the validity of these comparisons. Indeed we would be misled if we take the agreement as a confirmation of validity. Only by conditioning on characteristics *functionally independent* of treatment can we guarantee a comparison of apples to apples. The downside of requiring functional independence between Z and T, however, is that it cannot be tested by data. This echoes Pearl's emphasis (see Pearl 2000, p. 180) that probability calculus is not rich enough for handling Simpson's Paradox. Fortunately, when we use Z to infer intrinsic characteristics, rather than for direct conditioning, we can circumvent this problem.

To proceed, we first note that at the resolution level defined by intrinsic characteristics, C, ω^* is indistinguishable from any individual in $\Omega_C(\omega^*)$. We can then approximate the infinite-resolution estimand $(Y(\omega^*, 0), Y(\omega^*, 1))$ by averaging $(Y(\omega, 0), Y(\omega, 1))$ over $\Omega_C(\omega^*)$ to obtain the lower-resolution, *operational estimand* $P(Y(\omega, t) = 1 | \omega \in \Omega_C(\omega^*))$ for $t = 0, 1$. The inferential question, "How strict should we be?" becomes "How should we choose C?" To increase the quality of our approximate clones, we want C to be as rich a set of intrinsic characteristics as possible. But the price is a loss in our capacity to estimate the operational estimand from the observed data. For one, we may not observe all the components of C, and even when we do, there may not be any observed individual that satisfies $\omega \in \Omega_C(\omega^*)$; see Meng (2014). We therefore

want our operational estimand $P(Y(\omega, t) = 1 | \omega \in \Omega_C(\omega^*))$ to be as close as possible to our *ideal estimand*, $Y(\omega^*, t)$, and simultaneously to approximate our operational estimand sufficiently well by an *estimator*. To strike the right balance, we need to understand the *data resolution*.

2.3 The Estimation Question: Data Resolution

Once we rigorously define the concept of intrinsic characteristics, there is no ambiguity over what variables to condition on: *condition on as many intrinsic characteristics as possible*. The question then turns to, "What is possible?." This is ultimately a problem of inferring intrinsic characteristics from observed data. The observation process, which we now define, determines the highest possible resolution for our inference, that is, maximally what we can say about intrinsic similarity from apparent similarity.

Suppose our clinical trial contains data produced by n realized states $\{(\omega_i, t_i)\}_{i=1}^n$. Let $C(\omega_i, t_i)$ denote a realized state (not necessarily intrinsic) characteristic. However, the *potential outcomes* $C_1(\omega_i) \equiv C(\omega_i, 1)$ and $C_0(\omega_i) \equiv C(\omega_i, 0)$ are intrinsic characteristics of individual ω_i because these functions themselves are unaffected by the treatment assignment t_i. What is affected is which of these potential outcomes we are allowed to observe. Then $C(\omega_i, t_i)$ can be thought as being generated via

$$C(\omega_i, t_i) = t_i C_1(\omega_i) + (1 - t_i) C_0(\omega_i). \quad (2)$$

Clearly (2) is applicable whether C is the dependent variable Y or the third variable Z, as seen in (1).

To connect back to the common missing-data setup, we can view $Z(\omega_i, t_i)$ as the observed data and $(Z_0(\omega_i), Z_1(\omega_i))$ as the missing or augmented data. We want to condition on the intrinsic characteristics $(Z_0(\omega_i), Z_1(\omega_i))$, but this requires us to infer/predict it from $Z(\omega_i, t_i)$. By expressing $Z(\omega, t)$ in terms of $(Z_0(\omega), Z_1(\omega))$ via (2), we see that conditioning on the third variable leads to the comparison

$$P(Y = 1 | T = 1, Z = z) - P(Y = 1 | T = 0, Z = z), \quad (3)$$

which is *mathematically equivalent* to

$$P(Y = 1 | T = 1, Z_1 = z) - P(Y = 1 | T = 0, Z_0 = z). \quad (4)$$

Whereas (3) gives us the illusion of holding everything else constant other than the treatment assignment, the explicit differential subscripts in the higher-resolution expression (4) reveal that we are actually comparing apples and oranges unless $Z_0(\omega) = Z_1(\omega)$ for all $\omega \in \Omega$. If we all adopted the explicit notation in (4), we believe much of the current confusion could have been avoided.

Obviously intrinsic characteristics generating no heterogeneity in the observed data cannot possibly distinguish *any* individuals in our data. Therefore, since by (2) ω_i can affect the generation of Z only through $(Z_0(\omega_i), Z_1(\omega_i))$, we know that the choice $C(\omega_i) = (Z_0(\omega_i), Z_1(\omega_i))$, $R = 2$, is the upper limit for how rich we can make $C(\omega_i)$. Thus, we can take R to be 0, 1, or 2, but which one is optimal?

2.4 Resurrecting The Third Objective: Finding an Optimal Resolution

Angrist, Imbens, and Rubin (1996) proposed studying subpopulations defined by all possible pairs of (Z_0, Z_1), termed *principal strata* by Frangakis and Rubin (2002). In clinical trials with noncompliance, compliance can be thought of as a side effect of the treatment assignment (see Jin and Rubin 2008), and hence we should compare treatments conditional on compliance type, (Z_0, Z_1). Pearl (2011) also saw value in using principal strata to classify individuals. What he criticized is their use in defining the notions of "direct" and "indirect effect." But when our primary objective is making a treatment choice for ω^*, the opinion seems to be unanimous that conditioning on principal strata gives us a better look (relative to no conditioning) at how differences in treatment differ across individuals. Have we resolved Simpson's Paradox—is the answer always to choose $R = 2$ with $C = (Z_0, Z_1)$?

Expression (2) with $C = Z$ shows that unless additional assumptions are made, we may only infer Z_{0i} when $T_i = 0$ and Z_{1i} when $T_i = 1$. Therefore, the *overall data resolution*, R_{data}, is 0: no intrinsic characteristic is observed for *all* subjects. Thus to reach $R = 2$, we need additional assumptions, typically in the form of prior specifications. But as usual, bias resulting from prior misspecification may overwhelm the resolutional benefit gained from using $R = 2$ rather than $R = 0$ or 1. At the other end of the spectrum, we might be tempted to force $R = R_{\text{data}} = 0$ for ease of estimation, but this choice throws away valuable information because for each individual ω_i, we do observe one (and only one) component of $(Z_0(\omega_i), Z_1(\omega_i))$. From this we cannot say which individuals are alike at resolution $R = 2$, *but* we can say which are unlike: ω_i is unlike ω_j if $Z_0(\omega_j) \neq 0$ when $Z_0(\omega_i) = 0$. As we see in Section 3, to exploit this partial information we will need to take $R > 0$ even if $R_{\text{data}} = 0$.

Just as we must carefully balance bias and variance in selecting a model, so we must pivot between the ideal high-resolution estimand and feasible low-resolution estimators in choosing an operational estimand to net a better treatment decision. There is no such thing as a *correct* or *natural* choice of C. Box's quote "All models are wrong but some are useful" now becomes "All choices of C are wrong (except $C(\omega) = \omega$) but some are useful."

3. LET US ENJOY SOME FORGOTTEN OR FORBIDDEN FRUITS

When a treatment decision is needed for an individual ω^*, a key quantity of interest is $\theta(\omega^*) = \text{sign}\{Y_1(\omega^*) - Y_0(\omega^*)\}$, which indicates if treatment $T = 1$ is better than ($\theta = 1$), worse than ($\theta = -1$), or the same as treatment $T = 0$ ($\theta = 0$). The science of inference then is to choose a suitable population average to infer this individual-specific estimand, that is, to approximate the ideal (infinite-resolution) estimand $\theta(\omega^*)$ by the operational estimand (a lower-resolution average), $E[\theta(\omega) | \omega \in \Omega_C(\omega^*)]$, using the notation of Section 2.2. However, this lower-resolution average is itself not directly available since we never observe $(Y_1(\omega), Y_0(\omega))$ jointly. Nevertheless, for binary (Y_0, Y_1),

$$E[\theta(\omega) | S] = E(Y_1 | S) - E(Y_0 | S) \quad (5)$$

holds for any S for which both sides of (5) are defined. This allows us to define an operational estimand:

$$E[\theta(\omega)|\Omega_C(\omega^*)] = E(Y_1(\omega)|\omega \in \Omega_C(\omega^*)) - E(Y_0(\omega)|\omega \in \Omega_C(\omega^*)). \quad (6)$$

To estimate (6), we can then choose an estimator of the form

$$\tilde{\theta}(\tilde{C}_0, \tilde{C}_1) \equiv E(Y_1(\omega)|\omega \in \Omega_{\tilde{C}_1}(\omega^*)) - E(Y_0(\omega)|\omega \in \Omega_{\tilde{C}_0}(\omega^*)), \quad (7)$$

where $\tilde{C}_0, \tilde{C}_1$ can comprise only of intrinsic characteristics that are directly identifiable from data. To simplify our discussions, we have assumed in (7) that our samples are large enough that we can replace any sample average by the corresponding population average. Whereas we must choose the same C in defining our operational estimand to ensure a comparison of apples and apples, we are unaware of any *estimation* principle that would prevent us from using different $\tilde{C}_0$ and $\tilde{C}_1$ in (7). That is, $\tilde{C}_0$ and $\tilde{C}_1$ must both be subvectors of C, but they do not need to coincide. As long as our goal is correct, we can and *should* be as Machiavellian as possible in reaching it.

However, some readers might be puzzled or even disturbed by the idea of allowing different $\tilde{C}_0$ and $\tilde{C}_1$. What egregious hypocrites we are, accusing others of comparing apples and oranges, while our own prescription seems to advocate comparing apples to apricots! Have we warned others away from this forbidden fruit only to gorge on it ourselves? Of course not. What is forbidden is to mistake *apparent similarity* for intrinsic similarity, not the use of *apparent dis-similarity* to (better) estimate intrinsic similarity. We saw in Section 2.3 that if we force $\tilde{C}_0 = \tilde{C}_1$, then the maximal resolution of $\tilde{C}_0 = \tilde{C}_1$ (without further assumptions) is $R_{\text{data}} = 0$. However, when $\tilde{C}_0$ and $\tilde{C}_1$ can differ, we can achieve $\tilde{R}_0 > 0$, $\tilde{R}_1 > 0$, where $\tilde{R}_t$ is the resolution of $\tilde{C}_t$; for an estimator $\tilde{\theta}(\tilde{C}_0, \tilde{C}_1)$ with $\tilde{R}_0 \neq \tilde{R}_1$, we will denote its resolution as $\frac{1}{2}(\tilde{R}_0 + \tilde{R}_1)$, leaving the question of the best designation to further research. This intuitive reasoning makes it easy to grasp why such an estimator can be better (e.g., having smaller mean squared error (MSE)) than the one forcing $\tilde{C}_0 = \tilde{C}_1$.

In Section 4.2, we will provide a theoretical condition, the "1/2 Rule" (26), justifying the use of $\tilde{C}_0 \neq \tilde{C}_1$. *The reality* is that $\tilde{C}_{1-t}$ is missing from group $T = t$, so we cannot use it in predicting Y_t (hence causing a mismatch in (7)). But *can we pretend* that we omitted $\tilde{C}_{1-t}$ not because it was not available but because it was not helpful (hence rendering the mismatch irrelevant)? The 1/2 Rule tells us when this pretense passes: the predictive power of what is missing must be less than half the predictive power of what is observed. The criterion used by the 1/2 Rule is MSE. Hence, it permits trading bias (incurred from our pretense) for variance (reduced by having $\tilde{R}_0, \tilde{R}_1 > R_{\text{data}}$).

For concreteness, let $\tilde{C}_t = Z_t$, the post-treatment cholesterol under treatment t. If the success of treatment t, Y_t, depends mostly on how the patient's cholesterol changes with respect to that treatment and only somewhat on how the patient's cholesterol changes under the alternate treatment, then the 1/2 Rule is satisfied. For those wondering why a patient's cholesterol under the alternate treatment *might* matter, such data can capture aspects of the patient's health status at time of treatment, which affect the *side effect* of the alternate treatment but not the side effect of the treatment applied. The question then is to what extent those same aspects affect the *main effect* of the treatment applied. The 1/2 Rule aims to characterize when we can use the part of the data that is easy to use and ignore the part of the data that is "hard" to use—by hard we mean data whose use would require a full Bayesian model for (Y_0, Y_1, Z_0, Z_1) thereby inviting (potentially very) biased prior information.

3.1 Low-Resolution Estimand or Low-Resolution Estimator?

The conclusion that direct conditioning on the third variable, (3), is valid only when $Z_0(\omega) \equiv Z_1(\omega)$ does not imply the next best alternative is no conditioning at all, that is, to use as the operational estimand:

$$\theta_{R=0} \equiv P(Y_1 = 1) - P(Y_0 = 1) = P(Y = 1|T = 1) - P(Y = 1|T = 0), \quad (8)$$

where the estimand resolution is $R = 0$ because no intrinsic characteristics are used. Note (8) does not hold in general without assuming that T is independent of (Y_0, Y_1), as emphasized by Wasserman (2013, June 20, Blog). As argued before, even if it is not legitimate to condition on Z, conditioning on the two-component (and hence $R = 2$) intrinsic characteristic (Z_0, Z_1) gives us a valid operational estimand:

$$\theta_{R=2}(z_0, z_1) \equiv P(Y_1 = 1|Z_0 = z_0, Z_1 = z_1) - P(Y_0 = 1|Z_0 = z_0, Z_1 = z_1), \quad (9)$$

which is a better approximation to our ideal estimand $\theta(\omega^*)$ than is (8). The question now becomes whether we can find a good enough estimator of (9) to exploit its higher resolution or whether the higher cost of estimating (9), as compared to (8), represents too large an investment. Let $\tilde{R}$ denote the resolution of an estimator for $\theta_{R=2}$. From this perspective, we see that the assumption, $Z_0 = Z_1$, is really the most convenient condition for achieving $\tilde{R} = 2$ by reducing (9) to (3), which permits the simplest estimation procedure. Simplicity is always welcome in practice, but must be assessed against the possible invalidity of too strong a condition. The multiresolution framework reminds us that weaker conditions do exist, and that $Z_0 = Z_1$ is not the only assumption that can motivate us to use (9) instead of (8).

To simplify our discussion, let us assume that Z_t's are binary, as in Dr. Armistead's article (e.g., $Z = 0$ and $Z = 1$ indicate, respectively, whether the patient's post-treatment blood pressure remained low or became normal). If individual i is assigned to treatment, $T_i = 1$, observing $Z_i = 1$ allows us to infer $Z_{1i} = 1$, but not whether the individual belongs to subpopulation $\{(Z_0, Z_1) = (0, 1)\}$ or to subpopulation $\{(Z_0, Z_1) = (1, 1)\}$. Given this reality, we have two strategies:

(i) *Lower* the resolution, R, of our operational estimand.
(ii) *Estimate* our high-resolution estimand, (9), by using a lower-resolution *estimator,* $\tilde{R} < R$.

The consequence of either choice is *resolution bias*—a mismatch of desired versus adopted resolution. For (i), the resolution bias is incurred in the *decision phase* (when we use the operational estimand to make a decision), whereas for (ii), it is

incurred in the *inference phase*. Let $D_{\tilde{\theta}} \equiv D(\tilde{\theta})$ be a decision function taking *estimates*, $\tilde{\theta}$, of $\theta(\omega^*) = \text{sign}\{Y_1(\omega^*) - Y_0(\omega^*)\}$ as argument and let $L(D_{\theta(\omega^*)}, D')$ be the loss of decision D' when the optimal decision is $D_{\theta(\omega^*)}$. Adopting the notation in the previous section, the two types of resolution bias are

$$\text{Inference bias:} \quad \tilde{\theta}(\tilde{C}_0, \tilde{C}_1) - E[\theta(\omega)|\Omega_C(\omega^*)] \tag{10}$$

$$\text{Decision bias:} \quad L\left(D_{\theta(\omega^*)}, D_{E[\theta(\omega)|\Omega_C(\omega^*)]}\right). \tag{11}$$

The bias we ultimately hope to minimize is

$$\text{Realized bias:} \quad L\big(D_{\theta(\omega^*)}, D_{\tilde{\theta}(\tilde{C}_0,\tilde{C}_1)}\big). \tag{12}$$

One can think about minimizing (12) directly, of course, but this reduced formulation masks the pivotal role of C, the mechanism through which we actually can influence the realized bias. To conceptually connect the realized bias back to the more helpful I-bias and D-bias, we rewrite the realized bias as

$$\begin{aligned} L\big(D_{\theta(\omega^*)}, D_{\tilde{\theta}(\tilde{C}_0,\tilde{C}_1)}\big) &= L\left(D_{\theta(\omega^*)}, D_{E[\theta(\omega)|\Omega_C(\omega^*)]}\right) + \Delta_L \\ &\quad \cdot \mathbb{I}\{|\tilde{\theta}(\tilde{C}_0, \tilde{C}_1) - E[\theta(\omega)|\Omega_C(\omega^*)]| > \tau\} \\ &= \text{D-bias} + \text{Estimation Penalty} \\ &\quad \cdot \mathbb{I}\{\text{I-bias} > \text{Tolerance}\}. \end{aligned} \tag{13}$$

Δ_L is the change in loss if $D_{\tilde{\theta}(\tilde{C}_0,\tilde{C}_1)}$ were used instead of $D_{E[\theta(\omega)|\Omega_C(\omega^*)]}$ (the optimal decision being $D_{\theta(\omega^*)}$) and τ is the amount of I-bias needed for our actual decision to deviate from the intended decision (under the operational estimand), that is, $D_{\tilde{\theta}(\tilde{C}_0,\tilde{C}_1)} \neq D_{E[\theta(\omega)|\Omega_C(\omega^*)]}$. Identity (13) is conceptually powerful because it reveals the balancing role played by C, which appears in both terms—the choice of C must balance the error from using a nonideal estimand for decision making (11) and the estimation error for that nonideal estimand (10).

This choice between (i) and (ii) is reminiscent of the bias-variance tradeoff. However, the bias-variance tradeoff takes place entirely within the inference phase, whereas the tradeoff between I-bias and D-bias occurs *across* phases. The utility of this two phase setup is to remind us that for coarse decisions, for example, treatment 1 versus treatment 0, a large amount of I-bias can be incurred without changing our final decision. On the other hand, D-bias by definition alters our decision from the optimum. The tolerance term, τ, captures this asymmetry in how I-bias and D-bias enter into the realized bias, hence distinguishing it from the usual bias-variance tradeoff. Expression (13) is most useful for binary decisions as it employs a single penalty and tolerance term. When the decision space is richer, (13) may be rewritten to exhibit additional thresholds (with associated penalties). But the emphasis is the same: we may prefer making stronger assumptions to estimate a higher-resolution operational estimand using low-resolution data—knowing full well that this estimate will be biased—over settling for a low-resolution operational estimand. The latter, even though estimated with certainty, may yet be meaningless or misleading.

3.2 Simpson's Warning and the ID-Bias Tradeoff

We begin our investigation of the ID-bias tradeoff with strategy (ii). Whereas the success rate of treatment 1 is observed in the superpopulation $\{Z_1 = z_1\}$, the corresponding rates in its two subpopulations, $\{(Z_0, Z_1) = (0, z_1)\}$ and $\{(Z_0, Z_1) = (1, z_1)\}$, though desired, are not. Under the usual mean-squared loss, the best prediction of the desired $P(Y_1 = 1|Z_0, Z_1 = z_1)$, as a function of the random variable Z_0, is its expectation conditional on $Z_1 = z_1$, which is the superpopulation success rate $P(Y_1 = 1|Z_1 = z_1) = P(Y = 1|T = 1, Z = z_1)$. This is equivalent to choosing $C = (Z_0, Z_1)$ in (6) and $\tilde{C}_1 = Z_1$ in (7). Similarly for $T = 0$, we choose $\tilde{C}_0 = Z_0$; as discussed, to *estimate* conditioning on $\Omega_C(\omega^*)$ we can choose $\Omega_{\tilde{C}_0}(\omega^*) \neq \Omega_{\tilde{C}_1}(\omega^*)$.

Applying this reasoning, we can estimate the $R = 2$ estimand (9) by the $\tilde{R} = 1$ estimator

$$\begin{aligned} \tilde{\theta}^{\tilde{R}=1}_{R=2}(z_0, z_1) = {} & P(Y = 1|T = 1, Z = z_1) \\ & - P(Y = 1|T = 0, Z = z_0), \quad z_0, z_1 \in \{0, 1\}. \end{aligned} \tag{14}$$

The use of the tilde notation $\tilde{\theta}$ instead of the usual hat notation $\hat{\theta}$ is to remind us that even if there is no sampling error—Equation (14) is a population mean instead of sample average—we will still have errors caused by the discrepancy between $\tilde{R}$ and R. Also note that an implicit assumption here is that both values of $\{0, 1\}$ are observed for Z under $T = 1$ and $T = 0$; in general we assume the support of Z_t is invariant to t (which may be violated, such as when the treatment shifts all cholesterol levels upward).

Expression (14) says that to conclude that treatment 1 is superior to treatment 0 for all principal strata $\{(Z_0, Z_1) = (z_0, z_1)\}$, we need to consider *four* separate comparisons. For the two comparisons with $z_0 = z_1$, our estimate $\tilde{\theta}^{\tilde{R}=1}_{R=2}(z_0, z_1)$ corresponds to exactly the two Z-conditional contrasts, (3). Thus, the Z-conditional contrasts have forgotten about the two principal strata with $Z_0 \neq Z_1$, where individuals observed to be dissimilar in the two treatment groups may be actually intrinsically similar (though we must be mindful of our I-bias in making this assertion). Of course, if all four comparisons share the same sign, then we pay no price for our forgetfulness (if we base our treatment decision only on the sign). This is precisely the situation where Simpson's Paradox does not occur. If all four comparisons implied by (14) are positive, we obtain $P(Y = 1|T = 1) > P(Y = 1|T = 0)$. Hence, either the Z-conditional contrasts or the marginal contrast, taken as a reduction of (14), preserves the full sign information contained in (14).

When Simpson's Paradox occurs, the sign of the estimated treatment effect on the subpopulation where $Z_0 = Z_1$ must differ from the sign of the estimated treatment effect over the subpopulation where $Z_0 \neq Z_1$. The Z-conditional contrasts contain the sign information for the subpopulation where $Z_0 = Z_1$, and the marginal contrast contains the sign information for the subpopulation where $Z_0 \neq Z_1$. Neither tells the entire story. Simpson's Paradox is not paradoxical at all from this viewpoint: the sign of the Z-conditional contrast, (3), and the sign of the marginal contrast, (8), *do not contain contradictory information, but rather orthogonal information* pertaining to *disjoint* subpopulations. Or as Dr. Armistead put it "an apparent contradiction that may contain more than one truth."

Ironically, the advice to use only the marginal contrast when we cannot assume $Z_0 = Z_1$ makes the same mistake as the

advice to use the Z-conditional contrast. They both throw away the sign information for parts of the population. Conceptually, we fell into the trap of thinking that either (8) or (3) must be correct, when they are both correct or incorrect, *depending on which individuals* the information will be applied to. The appearance of Simpson's Paradox provides evidence for *treatment effect by subpopulation interaction*. A low-resolution estimand, for example, (8), will incur high D-bias, because the optimal decision may differ across principal strata. Instead of a paradox, the lesson we are given is *Simpson's Warning*:

> *Low resolution operational estimands are most dangerous (higher D-bias) when Simpson's Paradox appears.*

Rather than telling us to default to a marginal comparison when $Z_0 \neq Z_1$, the appearance of Simpson's Paradox is a sign that we should consider taking on I-bias to make a high-resolution inference, for example, (14), accounting for treatment effect by subpopulation interactions. However, in any particular situation, one may still feel that the I-bias incurred from using $\tilde{\theta}^{\tilde{R}=1}_{R=2}$ to estimate $\theta_{R=2}$ trumps any reduction in D-bias.

To make a more satisfactory ID-bias tradeoff, we can choose an operational estimand with resolution between $\theta_{R=0}$ and $\theta_{R=2}$. We can form *marginal principal strata* defined by Z_0 and Z_1 individually instead of jointly. This leads to operational estimands at resolution $R = 1$:

$$\theta_{R=1,Z_t}(z) = P(Y_1 = 1|Z_t = z) - P(Y_0 = 1|Z_t = z), \quad (15)$$

for $t = 0, 1$. When our target is $\theta_{R=1,Z_0}$, because Z_{0i} is observed for everyone assigned to treatment 0, we can estimate the $P(Y_0 = 1|Z_0 = z)$ term in (15) with no I-bias by setting $\tilde{C}_0 = Z_0$. For individuals in the treatment 1 group, we observe Z_{1i} but not Z_{0i}. In the absence of prior knowledge of the correlation between Z_{0i} and Z_{1i}, we once again estimate subpopulation rates using superpopulation rates, that is, we set $\tilde{C}_1$ to be empty in (7). The price of this strategy is I-bias but by lowering the resolution of our estimand to $R = 1$, we need only pay this price for the $P(Y_1 = 1|Z_0 = z)$ term in (15) (though logically it is possible for the difference of two biased estimators to be unbiased for the difference of their estimands). Our estimator then is

$$\tilde{\theta}^{\tilde{R}=0.5}_{R=1,Z_0}(z) = P(Y = 1|T = 1) - P(Y = 1|T = 0, Z_0 = z), \quad (16)$$

with resolution $\tilde{R} = (\tilde{R}_0 + \tilde{R}_1)/2 = 0.5$. In contrast to (14), which requires four comparisons, (16) requires only two comparisons, reflecting the increased D-bias caused by lowering our resolution. Specifically, at resolution $R = 1$, we take into account how the treatment effect may change across subpopulations defined by Z_0 $(= 0, 1)$ but ignore further changes in the treatment effect within those subpopulations.

In a nutshell, the resolution framework urges us to carefully consider the risk of adopting subpopulation specific *but I-biased* estimators versus robust *but D-biased* estimators. In addition, the resolution framework reminds us that $\theta_{R=0}$ and $\theta_{R=2}$ are not the only possible ways to compare the treatments.

3.3 Increase Data Resolution via Eliminating Subpopulations

In practice, some combination of lowering the estimand resolution and increasing the data resolution may be necessary to achieve a satisfactory ID-bias tradeoff. First, we note that the strong assumption $Z_0 = Z_1$ increases the estimation resolution to $\tilde{R} = 2$ by ruling out two subpopulations: $\{(Z_0, Z_1) = (0, 1)\}$ and $\{(Z_0, Z_1) = (1, 0)\}$. This assumption reduces the dimension of our estimand to make it identified. But we can weaken this assumption by ruling out only a single subpopulation, achieving estimators with resolution $\tilde{R} = 1.5$. We will incur I-bias in estimating $\theta_{R=2}$ but gain some robustness. Specifically, we can make the following "no-defier" assumption (see Angrist, Imbens, and Rubin 1996).

Exclusion Assumption. The subpopulation defined by $\{(Z_0, Z_1) = (1, 0)\}$ is empty.

In the context of blood pressure, this assumption says that treatment 1 performs at least as well as treatment 0 in raising the patient's blood pressure. A plausible scientific story is the existence of an unobserved genetic factor G, where $G = 0, 1, 2$ represent, respectively, the homozygote recessive, the heterozygote, and the homozygote dominant individuals. The homozygotes recessive and dominant are disposed toward low and normal blood pressure, respectively, regardless of treatment. The heterozygote is not predisposed toward either low or normal blood pressure—in this case, blood pressure is decided by treatment rather than genetic causes. Specifically, we have

$$Z = \begin{cases} 0 \text{ if } G = 0 \\ T \text{ if } G = 1 \\ 1 \text{ if } G = 2 \end{cases} \Rightarrow \begin{array}{l} \{(Z_0, Z_1) = (0, 0)\} = \{G = 0\} \\ \{(Z_0, Z_1) = (0, 1)\} = \{G = 1\}. \\ \{(Z_0, Z_1) = (1, 1)\} = \{G = 2\} \end{array}$$

Consequently, conditioning on principal strata allows us to successfully condition on the appropriate genetic factor even though it is unobserved and possibly even unknown to us. The nonexistence of the subpopulation $\{(Z_0, Z_1) = (1, 0)\}$ is induced by the ternary nature of genotypes. Whereas this story is useful for explaining the intuition behind principal strata, our calculations below do not depend on it.

As Figure 1 shows, by eliminating one principal stratum, we can directly infer (Z_{0i}, Z_{1i}) from T_i and Z_i whenever $T_i \neq Z_i$. Hence, the exclusion assumption increases the resolution of our

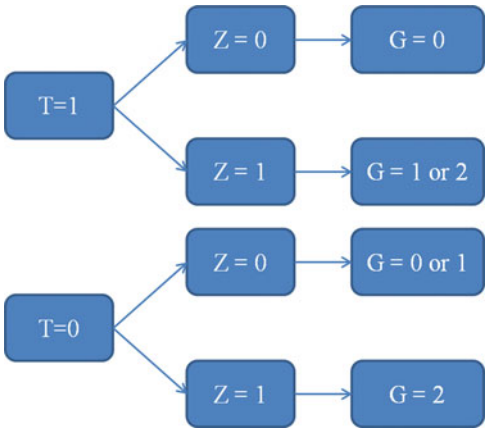

Figure 1. Inferring membership in principal strata from observed data.

Table 1. Lindley–Novick dataset

$Z=0$ (Low BP)	$Y=0$	$Y=1$	n	$Z=1$ (Normal BP)	$Y=0$	$Y=1$	n
$T=0$	21	9	30	$T=0$	3	7	10
$T=1$	8	2	10	$T=1$	12	18	30
Totals	29	11	40	Totals	15	25	40

data. But even for those individuals with $T_i = Z_i$, we still have some information on their likely stratum membership. To see this, denote $p_{ij} = P(Z_0 = i, Z_1 = j)$ for $i, j = 0, 1$. Then the exclusion assumption implies that

$$p_{00} = P(Z_1 = 0) = P(Z = 0|T = 1) \text{ and}$$
$$p_{11} = P(Z_0 = 1) = P(Z = 1|T = 0). \quad (17)$$

This allows us to estimate $\{p_{ij}, i, j = 0, 1\}$ directly from the data, because $p_{10} = 0$ and $p_{01} = 1 - p_{00} - p_{11}$.

Next we see that under the exclusion assumption we need to make only three comparisons:

$$\theta_{R=2}(g) = P(Y_1 = 1|G = g) - P(Y_0 = 1|G = g) \equiv \pi_{1g} - \pi_{0g}, \ g = 0, 1, 2. \quad (18)$$

To estimate $\{\pi_{tg}\}$ for $t = 0, 1$ and $g = 0, 1, 2$, let $\mu_{t,z} = P(Y = 1|T = t, Z = z)$. We then have

$$\mu_{1,0} = \pi_{10}, \quad \mu_{0,1} = \pi_{02}; \quad (19)$$
$$\mu_{0,0} = \frac{p_{00}}{p_{00} + p_{01}} \cdot \pi_{00} + \frac{p_{01}}{p_{00} + p_{01}} \cdot \pi_{01},$$
$$\mu_{1,1} = \frac{p_{01}}{p_{01} + p_{11}} \cdot \pi_{11} + \frac{p_{11}}{p_{01} + p_{11}} \cdot \pi_{12}. \quad (20)$$

Because all four $\mu_{t,z}$'s are directly estimable from the observed data (assuming we have observations with both values of Z under either treatment), π_{10} and π_{02} are in turn directly estimable because of (19), but π_{00}, π_{01}, π_{11}, and π_{12} are not because the equations in (20) are under-determined. Thus, our estimator will have resolution $\tilde{R} < 2$ and incur I-bias in estimating $\theta_{R=2}$.

However, for making a treatment decision, we may only need to infer sign($\theta_{R=2}$). Therefore, even if we do not know each (π_{0g}, π_{1g}) exactly, we may still able to determine whether or not $\pi_{1g} > \pi_{0g}$, eliminating I-bias for estimating sign($\theta_{R=2}$). For example, letting $o_{1|01} = p_{01}/p_{00}$ (odds of $G = 1$ given $G = 0$ or 1) and $o_{1|12} = p_{01}/p_{11}$ (odds of $G = 1$ given $G = 1$ or 2) yields, respectively,

$$\pi_{00} = (1 + o_{1|01})\mu_{0,0} - o_{1|01}\pi_{01},$$
$$\pi_{12} = (1 + o_{1|12})\mu_{1,1} - o_{1|12}\pi_{11}. \quad (21)$$

Using the fact that all π_{tg}'s must stay inside [0, 1], the first equation above restricts π_{00} to the interval

$$\max\{0, \mu_{0,0}(1 + o_{1|01}) - o_{1|01}\} \le \pi_{00} \le \min\{1, (1 + o_{1|01})\mu_{0,0}\}, \quad (22)$$

and similarly we can derive bounds for π_{12}. Such bounds, if sufficiently tight, allow for estimation of sign($\theta_{R=2}$) even when $\theta_{R=2}$ is not directly estimable. Furthermore, even if the bounds do not lead to a definite estimate of sign($\theta_{R=2}$), they may still enable us to make essentially bias-free decisions at a higher-resolution level with the help of extremely weak prior information, as we demonstrate below.

Example 2. Table 1 gives the Lindley–Novick dataset, as modified in Dr. Armistead's article. For simplicity, we will treat it as the population of interest instead of merely a sample and hence we can ignore any hat notation (but retain the tilde notation as needed). As noted in the article, as a *side effect*, treatment 1 appears to raise blood pressures more than treatment 0 does, so the exclusion assumption is not contradicted by the data. Adopting this assumption, by (17) we obtain that $p_{00} = p_{11} = 10/40 = 0.25$ and hence $p_{01} = 0.5$. That is, half the population comprises of individuals for whom treatment 1 was more effective than treatment 0 in increasing the blood pressure and the other half have blood pressure invariant to treatment choice. Equations (19) and (20) now become

$$0.2 = \mu_{1,0} = \pi_{10}, \quad 0.7 = \mu_{0,1} = \pi_{02};$$
$$0.3 = \mu_{0,0} = \frac{1}{3}\pi_{00} + \frac{2}{3} \cdot \pi_{01}, \quad 0.6 = \mu_{1,1} = \frac{2}{3}\pi_{11} + \frac{1}{3}\pi_{12}. \quad (23)$$

Using (23), we can derive bounds for the remaining success rates:

$$0 \le \pi_{00} \le 0.9, \quad 0 \le \pi_{01} \le 0.45,$$
$$0.4 \le \pi_{11} \le 0.9, \quad 0 \le \pi_{12} \le 1. \quad (24)$$

These bounds are loose. Hence, we are unable to directly conclude from them a definite sign for $\theta_{R=2}(g) = \pi_{1g} - \pi_{0g}$. (Section 4.1 will provide a direct link of this phenomenon to Simpson's Paradox.)

The closest we come to a definitive conclusion is in the subpopulation $\{(Z_0, Z_1) = (0, 1)\}$, where we can conclude that $\theta_{R=2}(1) \ge -0.05$, which suggests rather strong evidence that it is more likely than not that $\theta_{R=2}(1) \ge 0$. As a matter of fact, from (23), we see that $\theta_{R=2}(1) < 0$ if and only if $\pi_{12} - \pi_{00} > 0.9$, which requires $\pi_{12} > 90\%$ and $\pi_{00} < 10\%$. However, if this were a real life application, doctors can usually ballpark the magnitude of the success rate for various treatments. What is unknown is a more fine-scale comparison between alternatives. Yet this weak prior information (e.g., common sense) alone may be sufficient to rule out extreme rates such as $\pi_{12} > 90\%$ or $\pi_{00} < 10\%$ as well as their opposite nature. Putting all the pieces together, this analysis says that we cannot be sure which treatment is better for those whose blood pressure will be equally affected by both treatments. However, there is rather strong evidence that the main advantage of treatment 1 over treatment 0,

if it exists, is through the superior effect of treatment 1 in raising blood pressure, echoing Dr. Armistead's finding.

3.4 What Fruits Are Worth the I-Bias Price?

When bounds of the form (22) are insufficient for estimating sign($\theta_{R=2}$) and no useful prior information is available, we can resort, as before, to estimating subpopulation rates using their superpopulation counterparts. For the $T = 0$ group, we can use the superpopulation $\{Z_0 = 0\}$ to cover the estimations for $\{Z_0 = 0, Z_1 = 0\}$ and $\{Z_0 = 0, Z_1 = 1\}$, that is, we set $\tilde{C}_0 = Z_0$ in (7) when estimating π_{00} and π_{01}. Similarly, we set $\tilde{C}_1 = Z_1$ when estimating π_{11} and π_{12}. For cases where a subpopulation is directly observed, we set $\tilde{C}_0 = \tilde{C}_1 = (Z_0, Z_1)$. As in (14), the price of this estimation is I-bias.

Example 2 (continued). Adopting the strategy above, we obtain

$$0.3 = \mu_{0,0} = \tilde{\pi}_{00} = \tilde{\pi}_{01}, \qquad 0.6 = \mu_{1,1} = \tilde{\pi}_{11} = \tilde{\pi}_{12}.$$

We can then use them to construct estimates for $\theta_{R=2}$ (note π_{10} and π_{02} are directly available at resolution 2 under the exclusion assumption and hence they do not "wear" tilde):

$$\tilde{\theta}_{R=2}^{\tilde{R}=1.5}(0,0) = \pi_{10} - \tilde{\pi}_{00} = -0.1,$$
$$\tilde{\theta}_{R=2}^{\tilde{R}=1}(0,1) = \tilde{\pi}_{11} - \tilde{\pi}_{01} = 0.3,$$
$$\tilde{\theta}_{R=2}^{\tilde{R}=1.5}(1,1) = \tilde{\pi}_{12} - \pi_{02} = -0.1.$$

Here, the estimator resolution for the principal stratum $\{Z_0 = 0, Z_1 = 0\}$ is $\tilde{R} = 1.5$ since $\tilde{C}_1 = C = (Z_0, Z_1)$ ($\tilde{R}_1 = 2$) and $\tilde{C}_0 = Z_0$ ($\tilde{R}_0 = 1$). A similar logic gives the estimator resolution for the principal stratum $\{Z_0 = 1, Z_1 = 1\}$. The stratum $\{Z_0 = 0, Z_1 = 1\}$ has a lower estimator resolution, $\tilde{R} = 1$, because both terms must be estimated: $\tilde{C}_0 = Z_0$ ($\tilde{R}_0 = 1$) and $\tilde{C}_1 = Z_1$ ($\tilde{R}_1 = 1$).

Our estimates reiterate that treatment 1 outperforms treatment 0 only when it does a better job in raising blood pressure. In fact, for individuals whose blood pressure is invariant to T, *statistically speaking*, treatment 1 actually fares worse. Under our exclusion assumption, 50% of individuals have blood pressure that is invariant to treatment choice. (This does not mean that treatment 1 is the wrong choice for 50% of new patients—a conclusion that relies on the fallacy of the electoral college: winning a majority in subpopulations comprising a majority of the superpopulation does not imply winning a majority of the superpopulation.) A marginal comparison hides this information. While the low-resolution operational estimand, $\theta_{R=0} = (20/40) - (16/40) = 0.1$, is robust, it is misleading because it ignores treatment effect by subpopulation interactions. By characterizing these interactions, we see that the doctor should try hard to ascertain whether the patient's blood pressure will be invariant to treatment choice, and make a decision accordingly. For example, if the genotype G truly regulates blood pressure according to our story, a doctor could ascertain $(Z_0(\omega^*), Z_1(\omega^*))$ prior to making a treatment decision through genetic screening (but this requires us to at least suspect the genetic effect, perhaps through a secondary study).

But what if we truly have no predictive accuracy for the value of $(Z_0(\omega^*), Z_1(\omega^*))$? Let us consider a case where $\theta_{R=0} > 0$ but we know nothing about the value of $Z_0(\omega^*)$ or $Z_1(\omega^*)$ except that they are equal (e.g., we cannot predict a patient's blood pressure after either treatment, but we have good reasons to believe that the impact of the treatment on the blood pressure, as a side effect, will be very similar). *If we must select a treatment*, the best decision—in terms of minimizing the probability of mistake—is to assign ω^* to treatment 1, which is wholly based on $\theta_{R=0}$. However, $\theta_{R=2}$ is not useless—the higher-resolution information tells us about the quality and risk of the decision based on $\theta_{R=0}$. In particular, if $\theta_{R=0}(0,0) > 0$ and $\theta_{R=0}(1,1) > 0$, then we know our decision based on $\theta_{R=0}$ is reliable in that it is invariant to any new information about (Z_0, Z_1). But if $\theta_{R=0}(0,0)$ and $\theta_{R=1}(1,1)$ are of opposite sign, then our decision will not be invariant to information on (Z_0, Z_1)—the lack of invariance measures the inadequacy of low-resolution information. This same logic generalizes to the case where $Z_0 \neq Z_1$. In our example, a decision based on $\theta_{R=0}$ will be invariant to new information on (Z_0, Z_1) only 50% of the time, that is, it is no more reliable than flipping a fair coin.

When the decision process is no longer binary but includes the option "gather more information," information on the risk of choosing treatment 1 (or 0) can be used directly in decision making. Since intrinsic characteristics are functionally independent of treatment, conceptually nothing prevents us from assessing (Z_0, Z_1) prior to treatment; after all, the treatment process itself is a particular measurement process, which uses T to tease out (Z_0, Z_1)—a process that always creates missing data depending on T. We can, for example, use a patient's medical history or genetic screening to predict $(Z_0(\omega^*), Z_1(\omega^*))$. That is, an estimate of $\theta_{R=2}$ can lead to a different decision even when $(Z_0(\omega^*), Z_1(\omega^*))$ is unknown, if the former makes us suspect that a treatment decision based on $\theta_{R=0}$ alone is of unacceptable quality.

Example 2 (continued). We explicitly quantify how the additional information in our estimate of $\theta_{R=2}$ translates into tighter bounds for the error probability of a decision based on $\theta_{R=0}$. For any subpopulation S, define $q_{ij}^S = P(Y_0 = i, Y_1 = j|\omega \in S)$ for $i, j = 0, 1$. Let $q_{\cdot 1}^S = q_{01}^S + q_{11}^S$ and $q_{1\cdot}^S = q_{10}^S + q_{11}^S$. Define the error rate of choosing treatment 1 for subpopulation S as

$$\varepsilon^S(1) \equiv q_{10}^S / \left(q_{10}^S + q_{01}^S\right).$$

This gives the probability that we will make the *incorrect* decision when our choice leads to different outcomes, that is, when it matters. Suppose that we wish to bound $\varepsilon^S(1)$ given $\theta^S \equiv q_{01}^S - q_{10}^S$, for a subpopulation, S; note θ^S is directly observable from data because

$$P(Y_0 = 0, Y_1 = 1|\omega \in S) - P(Y_0 = 1, Y_1 = 0|\omega \in S)$$
$$= P(Y_1 = 1|\omega \in S) - P(Y_0 = 1|\omega \in S). \qquad (25)$$

Then we find the maximum and minimum (over q_{10}^S) of $\varepsilon^S(1) = q_{10}^S/(2q_{10}^S + \theta^S)$. For the Lindley–Novick data, we want the error rate over the entire population since nothing is known about our patient. We first estimate this rate using only information at resolution $R = 0$. We observe $\theta = \theta_{R=0} = 0.1$. In addition,

$0 \le q_{10} \le q_{1\cdot}$, where $q_{1\cdot}$ can be estimated by $P(Y_0 = 1) = 0.4$. This leads to bounds $0 \le \varepsilon(1) \le 4/9$.

Moving to higher-resolution (but estimated with I-bias) information at resolution $R = 2$, we see that since $\tilde{\theta}^{\tilde{R}=1.5}_{R=2}(0, 0) = -0.1$, (25) implies that $0.1 \le q^{(0,0)}_{10}$, where S here is the subpopulation $\{Z_0 = 0, Z_1 = 0\}$. Similarly $\tilde{\theta}^{\tilde{R}=1.5}_{R=2}(1, 1) = -0.1$ implies $0.1 \le q^{(1,1)}_{10}$. The lower bound for $q^{(0,1)}_{10}$ cannot be improved from 0 since $\tilde{\theta}^{\tilde{R}=1.5}_{R=2}(0, 1) = 0.3$. Using these bounds together with the fact $p_{00} = p_{11} = 0.25$ and $p_{01} = 0.5$ yields an improved lower bound for $q_{10} = \sum_{i,j} q^{(i,j)}_{10} p_{ij} \ge 0.05$. We then minimize and maximize $\varepsilon(1) = q_{10}/(2q_{10} + \theta)$ over $0.05 \le q_{10} \le 0.4$ for $\theta = 0.1$ to obtain $1/4 \le \varepsilon(1) \le 4/9$.

By incorporating higher-resolution information, we substantially improve the lower bound for $\varepsilon(1)$—in particular, the bounds using only low-resolution information are too optimistic about treatment 1. In the absence of information about $(Z_0(\omega^*), Z_1(\omega^*))$ for our new patient, our best decision (when we cannot "gather more information") is still to choose treatment 1, that is, we still base our decision on $\theta_{R=0}$. But the higher-resolution information allows us to ascertain the uncertainty of our decision. A fitting analogy to classical statistics is the difference between a point estimate and a confidence interval. (It is doubly fitting because the uncertainty ascertainment itself is subject to error: it relies on lower-resolution estimators such as $\tilde{\theta}^{\tilde{R}=1.5}_{R=2}$, just as we typically estimate the variance term when constructing a confidence interval.)

3.5 A Compromising Resolution Without Compromising Inference

When we use I-biased estimates, as in the previous section, we obviously should worry about the sensitivity of our inferences to the I-bias incurred. One way to ascertain this sensitivity is to lower the resolution of our operational estimand—hence lowering the I-bias. We can do so by focusing on $\theta_{R=1,Z_0}$ rather than on $\theta_{R=2}$. The marginal stratum $\{Z_0 = 0\}$ is a mix of the principal strata $\{(Z_0, Z_1) = (0, 0)\}$ and $\{(Z_0, Z_1) = (0, 1)\}$, equivalently $\{G = 0\}$ and $\{G = 1\}$, whereas $\{Z_0 = 1\} = \{G = 2\}$. We have lowered the I-bias because the first equation in (20) becomes $\mu_{0,0} = P(Y_0 = 1|Z_0 = 0)$, permitting a direct estimation of the control success rate in $\{Z_0 = 0\}$. We no longer need bounds for π_{00} and π_{01}, effectively reducing the undetermined parameters to just π_{11} and π_{12} in (20). This reduction is especially powerful if π_{00} or π_{01} were the quantities that we could not sufficiently bound. If it was π_{11} or π_{12} in (20) for which practical bounds did not exist, then we should consider using $\theta_{R=1,Z_1}$ instead of $\theta_{R=1,Z_0}$. If neither equation in (20) provides useful bounds, then nonnegligible I-bias accrues even at resolution $R = 1$.

Example 2 (continued). Let $\varphi_{tz} = P(Y_t = 1|Z_0 = z)$. Then

$$\varphi_{10} = \frac{1}{3}\pi_{10} + \frac{2}{3}\pi_{11}, \qquad \varphi_{00} = 0.3,$$
$$\varphi_{11} = \pi_{12}, \qquad \varphi_{01} = 0.7.$$

Our inability to bound π_{12} in (24) means that φ_{11} remains unidentified. Thus, we cannot compare the two treatments for the marginal stratum $\{Z_0 = 1\}$ without incurring I-bias. However, since $\pi_{10} = \mu_{1,0} = 0.2$ and $0.4 \le \pi_{11} \le 0.9$ (from (24)), by substituting this information into the equation for φ_{10}, we obtain $1/3 \le \varphi_{10} \le 2/3$. This implies that $\varphi_{00} = 0.3 < \varphi_{10}$. Thus, *statistically speaking*, treatment 1 outperforms treatment 0 if a patient's blood pressure remains low under treatment 0. This conclusion is reached without I-bias and relies only on the exclusion assumption. In addition, we know that the population proportion of such individuals is $P(Z_0 = 0) = 0.75$. This allows us to say that even if Z_0 were known for our patient (and even if we somehow discovered the value of φ_{11}), there is at least a 75% chance that we would not change our decision from the one based on $\theta_{R=0}$. (Again, this differs from the incorrect assertion that treatment 1 will outperform treatment 0 with probability at least 75%.)

Therefore, as before, the higher-resolution information allows us to calibrate the reliability of a decision based on $\theta_{R=0}$—reliability in the sense of invariance of our decision to new sources of knowledge. This assessment of reliability comes without I-bias, but it does come with D-bias in that we can only speak about decision invariance to newfound information about Z_0, not the more refined (Z_0, Z_1). By lowering our resolution, we ignore that the treatment effect may differ in the subpopulations $\{Z_0 = 0, Z_1 = 0\}$ and $\{Z_0 = 0, Z_1 = 1\}$. The decrease in I-bias at the cost of D-bias explains why our conclusion here differs from our conclusion in Section 3.4 that there is only a 50% chance that our decision will be invariant to new knowledge. However, *both* inferences carry much more (and higher quality) information than if we had chosen $\theta_{R=0}$ as our operational estimand, that is, if we had simply compared $P(Y = 1|T = 1) = 0.5$ to $P(Y = 1|T = 0) = 0.4$ and concluded that treatment 1 is better on average. Information on the quality (and risk) of our decision carried in the higher-resolution inference but missing from the lower-resolution operational estimand can be quite valuable when designing long-term treatment plans in practice.

4. MORE FRUIT FOR THOUGHTS

4.1 A Warning and Also a Dilemma: Scylla or Charybdis?

The loose bounds for $(\pi_{00}, \pi_{01}, \pi_{11}, \pi_{12})$ in the Lindley–Novick dataset indicate that the data are ambiguous about the comparative treatment effectiveness for specific principal strata $\{(Z_0, Z_1) = (z_0, z_1)\}$. The slackness of these bounds turns out to be directly related to the presence of Simpson's Paradox. In fact, we have the following result:

> Suppose that Simpson's Paradox occurs in a dataset. Then we will not be able to bound $(\pi_{00}, \pi_{01}, \pi_{11}, \pi_{12})$ to guarantee $\pi_{1g} \ge \pi_{0g}$ for $g = 0, 1, 2$, even under the exclusion assumption.

To prove this, first note that by the law of total probability, $\pi_{1g} \ge \pi_{0g}$ for $g = 0, 1, 2$ implies $P(Y_1 = 1) \ge P(Y_0 = 1)$ and hence $P(Y = 1|T = 1) \ge P(Y = 1|T = 0)$ for a randomized experiment. We now show that this contains enough information for the following statement to also hold for $z = 0, 1$:

$$\mu_{1,z} \equiv P(Y = 1|T = 1, Z = z) \ge P(Y = 1|T = 0, Z = z) \equiv \mu_{0,z}.$$

For the bound $\pi_{00} \leq \pi_{10}$ to hold for all values of π_{01}, by the first expression in (21), we must have

$$\max_{\pi_{01}} \pi_{00} = \mu_{0,0}(1 + o_{1|01}) \leq \pi_{10} = \mu_{1,0},$$

which implies $\mu_{1,0} \geq \mu_{0,0}$. Similarly for the bound $\pi_{12} \geq \pi_{02}$ to hold for all values of π_{11}, by the second expression in (21), we require

$$\min_{\pi_{11}} \pi_{12} = \mu_{1,1}(1 + o_{1|12}) - o_{1|12} \geq \pi_{02} = \mu_{0,1}.$$

Rewriting this inequality gives

$$\mu_{1,1} \geq \mu_{0,1} + \frac{o_{1|12}}{1 + o_{1|12}}(1 - \mu_{0,1}) \geq \mu_{0,1}.$$

Hence, the marginal and Z-conditional contrasts share a common sign.

The result says that if Simpson's Paradox occurs, then the data, even with our exclusion assumption, do not contain sufficient evidence for the superiority of treatment 1 over treatment 0 across all nonempty subpopulations. This statement has two sides. First, we see again that Simpson's Paradox occurs precisely when the better treatment may differ across subpopulations. The presence of this interaction reduces the utility of low-resolution operational estimands such as $\theta_{R=0}$, which give us no information about the quality/risk of any decision we make. The D-bias of low-resolution estimands will be high. Is there a way to make a high-resolution inference in this case *without* incurring significant I-bias?

The second side of the statement answers this question in the negative. We cannot hope to make a decision at resolution $R = 2$ relying on only the bounds supplied by the exclusion assumption. We will either have to take on I-bias by using superpopulation averages to estimate subpopulation averages or lower the resolution of our operational estimand, say to $R = 1$. Thus, there is a genuine tradeoff between I-bias and D-bias to be made. At this juncture, we must make the decision whether to sail closer to Scylla or Charybdis. For the Lindley–Novick data, we were able to reach a healthy compromise by comparing two treatments within subpopulations defined by Z_0 alone (a resolution $R = 1$ inference). We found statistical evidence for treatment 1 to be superior for individuals with low blood pressure under treatment 0. This conclusion is free of I-bias and with less D-bias than the operational estimand $\theta_{R=0}$.

Simpson's Warning foreshadows not just the potential for D-bias but dashes any hopes for reducing D-bias without incurring some I-bias. We need to strike a resolution-robustness compromise to obtain the most *useful* and *reliable* decision. We feel that the current thinking leans too heavily in favor of minimizing I-bias: defaulting to the resolution zero estimand, $P(Y_1 = 1) - P(Y_0 = 1)$. This is a bad habit of statisticians in thinking only in terms of estimation error—we would rather estimate a bad model correctly than estimate a good one poorly. Certainly, we should take some advice from crafty Odysseus who opted to lose a few men to Scylla rather than his whole crew to Charybdis. In our minds, the D-bias of low-resolution estimands is Charybdis—how could we make the right decision if we are asking the wrong questions? The I-bias from trying to obtain a higher-resolution inference is the few men we must sacrifice to Scylla to save the decision-making enterprise as a whole.

4.2 The Nightshade is Actually A Tomato

In Section 3, we proposed estimators for $\theta = Y_1 - Y_0$ of the form (7), which maximize the resolution component-wise. That is, in treatment group $T = t$, we approximate the population $\Omega_C(\omega^*)$ by $\Omega_{\tilde{C}_t}(\omega^*) = \Omega_{C_t}(\omega^*)$ where C_t contains all intrinsic characteristics observed for that group. The alternative chooses identical approximating populations for both groups: $\Omega_{\tilde{C}_0}(\omega^*) = \Omega_{\tilde{C}_1}(\omega^*) = \Omega_{C_{\text{com}}}(\omega^*)$, conditioning only on those intrinsic characteristics observable in *both* treatment groups, leading to the estimator

$$\begin{aligned}\tilde{\theta}(C_{\text{com}}) \equiv\ & E[Y_1(\omega) \,|\omega \in \Omega_{C_{\text{com}}}(\omega^*)] \\ & - E[Y_0(\omega) \,|\omega \in \Omega_{C_{\text{com}}}(\omega^*)] = E[\theta|C_{\text{com}}].\end{aligned}$$

Note that $\Omega_{C_0}(\omega^*)$ and $\Omega_{C_1}(\omega^*)$ are both refinements of $\Omega_{C_{\text{com}}}(\omega^*)$. The possible superiority of the maximal component-wise resolution estimator, $\tilde{\theta}(C_0, C_1)$, over $\tilde{\theta}(C_{\text{com}})$ underpins the justification for choosing an operational estimand with resolution greater than the data resolution. Hence, the theory of such estimators will be a crucial component in the future development of multiresolution inference. Below, we offer some low-hanging but nonetheless rich fruit to hopefully entice others.

Our goal is to characterize when $\tilde{\theta}(C_0, C_1)$ dominates $\tilde{\theta}(C_{\text{com}})$ in MSE. In evaluating the frequentist properties of $\tilde{\theta}(C_0, C_1)$ and $\tilde{\theta}(C_{\text{com}})$, it makes sense to condition on C_{com}. To simplify notation, let $\hat{Y}_t \equiv E[Y_t(\omega)|C_t]$, $R_t \equiv Y_t - \hat{Y}_t$, $\sigma_t^2 \equiv V(\hat{Y}_t|C_{\text{com}})$, and for $t = 0, 1$,

$$\begin{aligned}\beta_{t|1-t}^{\text{obs}} &\equiv \frac{\text{cov}(\hat{Y}_t, \hat{Y}_{1-t}|C_{\text{com}})}{V(\hat{Y}_{1-t}|C_{\text{com}})} \equiv \frac{\sigma_{01}^{\text{obs}}}{\sigma_{1-t}^2}, \\ \beta_{t|1-t}^{\text{mis}} &\equiv \frac{\text{cov}(R_t, \hat{Y}_{1-t}|C_{\text{com}})}{V(\hat{Y}_{1-t}|C_{\text{com}})} \equiv \frac{\sigma_{t,1-t}^{\text{mis}}}{\sigma_{1-t}^2}.\end{aligned}$$

Then $\tilde{\theta}(C_0, C_1)$ attains smaller MSE than $\tilde{\theta}(C_{\text{com}})$ if and only if $\sigma_{1,0}^{\text{mis}} + \sigma_{0,1}^{\text{mis}} \leq \frac{1}{2}(\sigma_0^2 + \sigma_1^2) - \sigma_{01}^{\text{obs}}$. To help us interpret, consider the case $\sigma_0^2 > 0$, $\sigma_1^2 > 0$, which allows us to rewrite the condition in terms of the *1/2 Rule*:

$$\begin{aligned}&\frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}\beta_{1|0}^{\text{mis}} + \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2}\beta_{0|1}^{\text{mis}} \\ &\qquad \leq \frac{1}{2}\left[\frac{\sigma_0^2}{\sigma_0^2 + \sigma_1^2}\left(1 - \beta_{1|0}^{\text{obs}}\right) + \frac{\sigma_1^2}{\sigma_0^2 + \sigma_1^2}\left(1 - \beta_{0|1}^{\text{obs}}\right)\right]. \quad (26)\end{aligned}$$

Here, σ_t^2 is the amount of variation in Y_t explained by the observed data C_t. So the weight $\sigma_t^2/(\sigma_0^2 + \sigma_1^2)$ is the fraction of total explained variation attributable to C_t. But how much of this information is *unique* to C_t, that is, would we have done any worse if the situation had been reversed and we predicted Y_t using C_{1-t} instead of C_t? To answer this, we regress $\hat{Y}_t$ on a function of C_{1-t}, $\hat{Y}_{1-t}$. If $\hat{Y}_{1-t}$ is perfectly redundant for the information in C_t, then the regression coefficient $\beta_{t|1-t}^{\text{obs}}$ equals 1. So $1 - \beta_{t|1-t}^{\text{obs}}$ measures the predictive power for Y_t *unique* to the observed data, C_t. Hence, the right-hand side of (26) equals 1/2 times a weighted average of the predictive information unique to the observed data for Y_0 and Y_1, respectively. Similarly, we can calculate the predictive power for Y_t that is unique to the

missing data, C_{1-t}. Note that R_t is variation in Y_t unexplained by C_t. Hence, the regression coefficient of R_t on $\hat{Y}_{1-t}$ measures how much information in C_{1-t} is missing from C_t. The left-hand side of (26) is a weighted average of the predictive information unique to the missing data. Thus, condition (26) can be interpreted as

$$\text{Info Unique to Missing Data} \leq \frac{1}{2} \cdot \text{Info Unique to Observed Data.} \quad (27)$$

Consider the special case where $\Omega_{C_0}(\omega^*) = \Omega_{C_{\text{com}}}(\omega^*)$, that is, the resolutional improvement is concentrated in predicting Y_1. Then, we require $\beta^{\text{mis}}_{0|1} \leq \frac{1}{2}$, that is, the information missing from group $T = 0$ should not be highly predictive of the residuals $Y_0 - E(Y_0|C_{\text{com}})$.

The danger of choosing $\tilde{C}_0 \neq \tilde{C}_1$ in (26) is that the missing data may be highly influential—inferences that condition only on the observed data will then fail to adequately approximate the operational estimand that conditions on both the observed *and* missing data. However, if dependence of outcome on the missing data is weak—in fact less than 1/2 the dependence of outcome on the observed data—then we can essentially "ignore the error term" and still enjoy the resolutional benefits of choosing our operational estimand so that $R = \dim(C) > \dim(C_{\text{com}})$. To make this interpretation clearer, we illustrate (26) on a canonical example.

Example 3. For analytic tractability, both (C_{0i}, C_{1i}) and (Y_{0i}, Y_{1i}) are taken to be continuous, though the intuition flows back easily to the discrete case. Assume that (C_{0i}, C_{1i}) are standardized but correlated normal variates: $N_2(\mathbf{0}_2, (1-\rho)\mathbf{I}_2 + \rho\mathbf{1}_2\mathbf{1}_2^T)$, which affect the potential outcomes (Y_{0i}, Y_{1i}) via

$$Y_t = \mu_t + \alpha^{\text{obs}} C_t + \alpha^{\text{mis}} C_{1-t} + \varepsilon_t$$

for $t = 0, 1$, where ε_0 and ε_1 are iid $N(0, \tau^2)$ and independent of (C_{0i}, C_{1i}). In the treatment group, we observe only C_1 and in the control group we observe only C_0. Thus, C_{com} is empty. The question of interest: when is $E(Y_1|C_1) - E(Y_0|C_0)$ a better predictor of $Y_1 - Y_0$ (in MSE terms) than $E(Y_1) - E(Y_0)$? As suggested, the answer depends on how important the missing information, C_{1-t}, is compared with the observed information, C_t, in predicting Y_t, determined here by the values of $(\alpha^{\text{obs}}, \alpha^{\text{mis}})$.

The best (in MSE terms) predictor of Y_t using C_t and the prediction residual are

$$\hat{Y}_t = E(Y_t|C_t) = \mu_t + (\alpha^{\text{obs}} + \alpha^{\text{mis}}\rho)C_t;$$
$$R_t = Y_t - \hat{Y}_t = \alpha^{\text{mis}}(C_{1-t} - \rho C_t) + \varepsilon_t.$$

To write condition (26), we find

$$1 - \beta^{\text{obs}}_{t|1-t} = \frac{1-\rho^2}{1+\rho}, \qquad \beta^{\text{mis}}_{t|1-t} = \frac{1-\rho^2}{\alpha^{\text{obs}}/\alpha^{\text{mis}} + \rho}.$$

These regression coefficients measure the predictive information unique to the observed and missing data, respectively—the similarity in form is striking with $\alpha^{\text{obs}}/\alpha^{\text{mis}}$ replacing the constant 1 in the latter. When $\alpha^{\text{obs}} > 0, \alpha^{\text{mis}} > 0$, then $\beta^{\text{mis}}_{t|1-t}$ is a monotonic decreasing function of $\alpha^{\text{obs}}/\alpha^{\text{mis}}$—the ratio of the predictive strength of the observed data to that of the unobserved data. Since $\sigma_0 = \sigma_1$, (26) then has the form

$$\frac{1-\rho^2}{\alpha^{\text{obs}}/\alpha^{\text{mis}} + \rho} \leq \frac{1}{2}\frac{1-\rho^2}{1+\rho}.$$

When $\rho \geq 0$, this becomes $\alpha^{\text{obs}}/\alpha^{\text{mis}} \geq 2 + \rho$. So if $\rho = 0$, we need the information in the observed data C_t to be twice as important as the information in the missing data, C_{1-t}. In many applications this is not an unreasonable assumption. After all, we would expect the potential side effects experienced by our patient under treatment t to be much more predictive of the success/failure of treatment t than the potential side effects experienced under alternative treatments. In these situations maximizing component-wise resolution will lead to smaller MSE than adopting the data resolution, namely, using the unconditional comparison $E(Y_1) - E(Y_0)$ to predict $Y_1 - Y_0$.

The limitation of (26) is that it only applies when MSE adequately describes our actual loss. For real life examples, other loss functions may be more appropriate: in selecting treatment for a patient, our focus may be 0–1 loss (did we choose the right or wrong treatment?). Nevertheless, the 1/2 Rule establishes the possibility of improving our decision making by maximizing resolution component-wise. And while the magic number may differ from 1/2 for other losses, the intuition remains: the explanatory power of the observed data must trump that of the missing data. We hope this insight will encourage others to taste this previously forbidden fruit, which is more tomato (yes, a tomato is a fruit) than nightshade. We ought to stop now before we over serve our dessert after Dr. Armistead's main entree, which was rich with food for thought. We do, however, hope that our multiresolution fruit basket is large and inviting enough for readers who remain hungry for more—please help yourself!

REFERENCES

Angrist, J. D., Imbens, G. W., and Rubin, D. B. (1996), "Identification of Causal Effects Using Instrumental Variables," *Journal of the American Statistical Association*, 91, 444–455. [21,24]

Frangakis, C. E., and Rubin, D. B. (2002), "Principal Stratification in Causal Inference," *Biometrics*, 58, 21–29. [21]

Jin, H., and Rubin, D. B. (2008), "Principal Stratification for Causal Inference With Extended Partial Compliance," *Journal of the American Statistical Association*, 103, 101–111. [21]

Lindley, D. V., and Novick, M. R. (1981), "The Role of Exchangeability in Inference," *The Annals of Statistics*, 9, 45–58. [20]

Meng, X.-L. (2014), "A Trio of Inference Problems That Could Win You a Nobel Prize in Statistics (If You Help Fund It)," in *The Past, Present and Future of Statistical Science*, eds. X. Lin, D. L. Banks, C. Genest, G. Molenberghs, D. W. Scott, and J.-L. Wang, Boca Raton, FL: CRC Press, pp. 535–560. [18,20]

Pearl, J. (2000), *Causality: Models, Reasoning and Inference*, Cambridge: Cambridge University Press. [18,20]

——— (2011), "Principal Stratification a Goal or a Tool?," *The International Journal of Biostatistics*, 7, 1–13. [19,21]

Rubin, D. B. (2005), "Causal Inference Using Potential Outcomes," *Journal of the American Statistical Association*, 100, 322–331. [18]

Wasserman, L. (2013, June 20), "Simpson's Paradox Explained," *Blog: Normal Deviate*. Available at http://normaldeviate.wordpress.com/. [18,20,22]