Randomized Controlled Trials and Trial Emulation

Ashley I Naimi

September 2026

Contents

1 Randomization

Randomization is one of the most foundational tools we have at our disposal as quantitative scientists. It's used in theory and practice to solve some very specific problems. Thinking generally, randomization is a tool for "extraction." It can lift a phenomenon out of its natural context so we can examine it on its own terms (Figure 1). For example, randomizing people to aspirin vs. placebo severs all the reasons why a doctor might prescribe it or a patient might take it.

Panel A of Figure 1 illustrates the problem. The exposure sits embedded in a web of other variables, some causing it, some caused by it. When we simply observe who gets the exposure in the real world, we cannot tell how much of any association with the outcome is due to the exposure itself versus the many other forces that determined who was exposed.

Randomization solves this by replacing all of those determinants with a single exogenous randomization mechanism independent of all other variables in the system (e.g., coin flip, Panel B). Because the randomization mechanism is unrelated to anything else in the system, the exposure is no longer associated with the variables that used to cause it. Any association between the randomized green node and the outcome must therefore flow through the exposure itself. This is what confounding control means in structural terms: severing the arrows into the exposure so that the only open path(s) are causal ones from the exposure to the outcome.

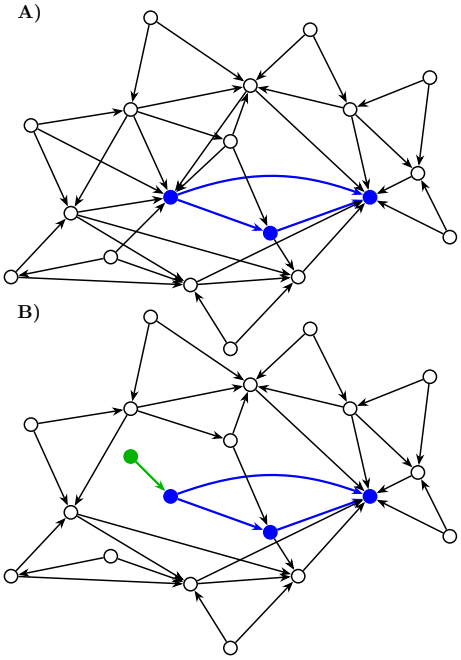

Figure 1: Complex network of variables, with an exposure–outcome effect of interest denoted in blue. Panel A shows the exposure–outcome effect embedded within a complex network of other variables. Panel B shows what happens when we randomly assign the exposure based on some exogenous green node.

2 Identifiability, Exchangeability, and Randomization

What does randomization give us mathematically? Let's follow the reasoning in Greenland and Robins (Greenland and Robins, 1986), which uses potential outcomes to think through the role of randomization.

What to read: Greenland & Robins Identifiability, Exchangeability, and Epidemiologic Confounding. *Int J Epidemiol* 1986;15(3):413-9

Potential outcomes are mathematical objects defined as outcomes that would have been observed under specific, possibly counterfactual, treatment/exposure scenarios. For example, Y^a is the outcome that would be observed if an individual were exposed to some treatment level $A = a$. At an individual-level, we can think of potential outcomes in terms of four response types:

	Y^1	Y^0	Interpretation
Type 1: Doomed.	1	1	Experiences the outcome whether exposed or unexposed.
Type 2: Causative.	1	0	Experiences the outcome only if exposed.
Type 3: Preventive.	0	1	Experiences the outcome only if unexposed.
Type 4: Immune.	0	0	Does not experience the outcome whether exposed or unexposed.

Suppose we observe a single treated person during a pre-defined risk period, who becomes diseased. We cannot tell which potential response type this person is just by looking at them, without further information.

Suppose further we observe a second untreated person during the same risk period. We now have four possible scenarios we could entertain: (a) both people fall ill, (b) only the treated person falls ill, (c) only the untreated person falls ill, (d) neither person falls ill. But even in this scenario, we can't know for sure if the exposure had an effect, since either one of these people could be any of the four potential response types.

For instance, the treated person may have been doomed, and the untreated person immune. In this case, it would look like our exposure had an effect, but this would not be true.

However, if we combine our data from these two individuals with an assumption that both individuals are of the same response type, we can make some progress in deducing whether the exposure had an effect. Scenarios (a) and (d), *combined with the assumption,* imply no exposure effect; Scenario (b) with the assumption implies a causative exposure effect occurred; and scenario (c) with the assumption implies a preventive exposure effect occurred.

We can only make progress in determining whether the exposure was causal when we combine our data on two persons (one treated, one untreated) with an assumption that the treated and untreated individuals are the *same* response type.

Now, continuing with the Greenland and Robins (1986) example, suppose we have data on a closed cohort of N_1 and N_0 treated and untreated individuals who are disease free at baseline. Suppose further that we let p_j and q_j, $j = 1, 2, 3, 4$ be the proportion of members in the N_1 and N_0 samples with response type j, respectively. From this, we can construct a two-by-two table:

We use this to define the risk of the disease in the treated and untreated as $R_1 = A_1/N_1$ and $R_0 = A_0/N_0$.

	Treated	Untreated
Cases	$A_1 = (p_1 + p_2)N_1$	$A_0 = (q_1 + q_3)N_0$
Non-cases	$B_1 = (p_3 + p_4)N_1$	$B_0 = (q_2 + q_4)N_0$

However, whether we had a situation where $R_1 - R_0 > 0$, $R_1 - R_0 < 0$, or $R_1 - R_0 = 0$, we can't tell whether the exposure is causal, protective, or has no effect because:

$$R_1 - R_0 = A_1/N_1 - A_0/N_0 \tag{1}$$

$$= \frac{(p_1 + p_2)N_1}{N_1} - \frac{(q_1 + q_3)N_0}{N_0} \tag{2}$$

$$= (p_1 + p_2) - (q_1 + q_3) \tag{3}$$

Based on this, it's possible that $R_1 - R_0 > 0$ because $p_1 > q_1$, while $p_2 = q_3$. In words, it's possible that the proportion of causative response types in the treated group (p_2, individuals who, if treated, experience the disease) is the same as the proportion of preventive response types in the untreated group (q_3, individuals who, if untreated, experience the disease), but that there are more doomed individuals in the treated group (p_1) than there are in the untreated group (q_1).

How do we avoid this problem? We need something that allows us to assume that the proportion of doomed individuals is the same in both treated and untreated groups. That way, we can assume that $p_1 = q_1$, and thus any comparison of R_1 and R_0 can be interpreted as a comparison of the proportion of causal and preventive response types.

Equivalence of these response type proportions ($p_1 = q_1$) can be thought of in terms of *exchangeability*: if the treatment status of the treated and untreated groups were *exchanged*, we'd still end up with the same proportion values for p_1 and q_1.

2.1 How Randomization Gives Exchangeability

Randomization gives us justification to assume equivalence between p_1 and q_1. To see how, consider the example in Chapter 2 of the Hernán and Robins book (Hernán and Robins, 2020, Chapter 2). Consider a setting in which we

are estimating the effect of aspirin on headache incidence in a cohort of individuals aged 18-40 years.[1] To do this experiment, a researcher randomly assigns 50% of the cohort to aspirin, and the remaining 50% to placebo. However, to overcome some logistical complications, before actually giving them aspirin/placebo, this researcher hands out cards that indicate whether the participant was assigned to aspirin (red card) versus placebo (blue card).

[1] Assume that our sample size is sufficiently large so as to avoid any sampling variability problems.

After the cards/aspirin/placebo are distributed and the follow-up period transpires, the researcher tallies up the number of headaches in each exposure group. He finds the following results:

$$\text{Aspirin (Red Card): } E(Y \mid X = 1) = 0.6$$
$$\text{Placebo (Blue Card): } E(Y \mid X = 0) = 0.1$$

However, after reviewing the study protocol, he realizes that he accidentally assigned placebo to those with the red card, and aspirin to those with the blue card, instead of the other way around. Fortunately, this has no actual impact on the study, with the exception of needing to switch the aspirin label with the placebo label. Why? Randomization (in a sufficiently large enough sample) creates independencies between outcome that would be observed under some exposure value (the potential outcome) and the observed exposure. In our case, $E(Y^{x=1}) = 0.1$, and this is the case whether the exposure received was placebo ($X = 0$) or aspirin ($X = 1$):

$$E(Y^{x=1}) = 0.1 \implies \begin{cases} E(Y^{x=1} \mid X = 1) = 0.1 \\ E(Y^{x=1} \mid X = 0) = 0.1 \end{cases}$$

Thus, because of randomization the following mathematical relation is implied:

$$E(Y^x \mid X) = E(Y^x) \tag{4}$$

which states that the potential outcomes are mean independent of the observed exposure.[2] Connecting this back to our example of potential response types, and the equivalence between p_1 and q_1, suppose that more doomed individuals were assigned to treatment than placebo. Connecting this back to our potential response types table, suppose that more doomed individuals were assigned to treatment than placebo. In this case, the observed exposure

[2] This is sometimes written in different ways, with subtlely different implications. See Hernán and Robins (2020), page 10-11.

X (being assigned to treatment) would be associated with the means of the potential outcomes. The mean of a potential outcome within an arm is nothing more than a sum of response type proportions.

The only response types with $Y^{x=1} = 1$ are the doomed and the causative, and the only response types with $Y^{x=0} = 1$ are the doomed and the preventive. Therefore:

$$E(Y^{x=1} \mid X = 1) = p_1 + p_2$$
$$E(Y^{x=1} \mid X = 0) = q_1 + q_2$$
$$E(Y^{x=0} \mid X = 1) = p_1 + p_3$$
$$E(Y^{x=0} \mid X = 0) = q_1 + q_3$$

Assume for a moment that the first and last of these quantities are observable: among those who took aspirin, the outcome we observe is the potential outcome we'd observe under "take aspirin", so the risk in the aspirin arm is $R_1 = p_1 + p_2$, and likewise the risk in the placebo arm is $R_0 = q_1 + q_3$.[3] The middle two quantities are counterfactual: $q_1 + q_2$ is the risk the placebo arm would have experienced had they taken aspirin, and $p_1 + p_3$ is the risk the aspirin arm would have experienced under placebo. With an excess of doomed individuals in the aspirin arm ($p_1 > q_1$, with the causative and preventive proportions equal across arms, i.e., no real effect), we would have:

[3] That the observed outcome among those who took aspirin is the outcome that would be observed under aspirin is itself an assumption, known as *consistency*, which we discuss in the section on identification.

$$E(Y^{x=1} \mid X = 1) = p_1 + p_2 > q_1 + q_2 = E(Y^{x=1} \mid X = 0)$$

and similarly for $Y^{x=0}$. The means of the potential outcomes would therefore differ across levels of the observed exposure. In our aspirin example, this would amount to the red-card group being more headache-prone than the blue-card group from the outset. The researcher's labeling mix-up would then have been a problem, because neither group's experience could stand in for the other's. In effect, the assignment mechanism cannot preferentially channel doomed individuals into one arm.

3 Randomized, Quasi-Randomized, and Non-Randomized Studies

Actual physical randomization is a tool that we use to justify the exchangeability assumption (in expectation). But there are approaches that we can use

in the absence of physical randomization to try justify the exchangeability assumption.

For example, in quasi-randomized studies, researchers rely on their understanding of some natural physical mechanism that enables them to believe that the treatment was allocated in a way that is close enough to random. For instance, Mendelian randomization relies on the fact that a given individual's germline genetic variants are inherited from their parents through the random segregation of alleles during meiosis. If the genes encoded in these alleles are highly predictive of some potential exposure variable of interest, one can reason that the exposure was "randomized" by the natural processes of meiosis. Researchers use methods for Mendelian randomization (i.e., instrumental variable methods, such as two-stage least squares estimators) to evaluate the relationship between an exposure and outcome of interest. For example, Timpson et al. (2009) used variants in the *FTO* (rs9939609) and *MC4R* (rs17782313) genes as instruments for BMI to try to estimate the causal effect of adiposity on blood pressure in a Danish population-based cohort. They found that a 10% increase in BMI caused a 3.85 mmHg increase in systolic blood pressure (95% CI: 1.88, 5.83) and a 1.79 mmHg increase in diastolic blood pressure (95% CI: 0.68, 2.90), providing evidence that greater adiposity elevates hypertension risk.

What to read:
Swanson et al Nature as a Trialist?: Deconstructing the Analogy Between Mendelian Randomization and Randomized Trials. *Epidemiol* 2017;28(5):653-659
Naimi AI and Whitcomb BW Defining and Identifying Local Average Treatment Effects *Am J Epidemiol* 2024;193(7):935-937

Note:
In the abstract of Timpson et al. (2009), they state "Avoiding the epidemiological problems of confounding, bias, and reverse causation, we confirmed observational associations between body mass index and blood pressure." This statement is misleading, would have been better had the comma between "confounding" and "bias" been removed. Why? Mendelian Randomization methods specifically (and instrumental variable methods more generally) are subject to important biases, but not necessarily confounding biases (assuming, of course, that the SNP is a valid instrumental variable, see Hemani et al. (2018)). Bias can indeed result from things like weak instruments (Burgess et al., 2011), population stratification (Sanderson et al., 2022), or pleiotropy (Davey Smith and Hemani, 2014; Hemani et al., 2018).

Other types of quasi-randomized studies rely on different "natural processes" to invoke "randomization" and thus exchangeability. For example, interrupted time-series analyses rely on the fact that a well-defined policy change or regulatory action creates a discrete "interruption" in an otherwise

continuous outcome trend over time (Turner et al., 2020). If the timing of this interruption is plausibly unrelated to pre-existing trends in the outcome (if the policy was implemented for reasons exogenous to the outcome trajectory) one can reason that exposure to the pre- versus post-policy period was allocated in a way that approximates randomization with respect to time. Researchers often use segmented regression methods to estimate changes in the level and slope of the outcome trend before and after the interruption point. For example, Orandi et al. (2023) used the 2011 FDA mandate limiting acetaminophen to 325 mg per tablet in prescription combination opioid products as an interruption to estimate its effect on hospitalization rates for acetaminophen-opioid toxicity in a national sample of US hospitalizations. They suggested that hospitalization rates fell from 12.2 per 100,000 hospitalizations (95% CI: 11.0, 13.4) before the mandate to 4.4 per 100,000 (95% CI: 4.1, 4.7) after, reflecting an absolute difference of 7.8 per 100,000 (95% CI: 6.6, 9.0) fewer hospitalizations, suggesting that the FDA mandate reduced acetaminophen-opioid toxicity hospitalizations.

What to read: Bernal et al Interrupted time series regression for the evaluation of public health interventions: a tutorial. *Int J Epidemiol*. 2016;46(1):348–355

Another quasi-randomized design is the regression discontinuity (RD) design. Similar to ITS, RD relies on the fact that many clinical and administrative decisions are governed by a discrete threshold applied to a continuous "running variable." These can include a clinical cutoff that triggers a treatment protocol. Individuals just below and just above this threshold are nearly identical in all other respects, so one can reason that treatment assignment in a narrow window around the cutoff was allocated in a way that approximates randomization. Researchers can use local linear regression, such as a kernel-weighted least squares estimator that down-weights observations farther from the cutoff, to fit separate regression lines on either side of the cutoff, and estimate the discontinuous "jump" in outcomes that occurs at the threshold. The magnitude of this "jump" is interpreted as the causal effect of treatment. For example, Almond et al. (2010) exploited the clinical 1,500 g birth weight threshold to estimate the causal effect of intensive neonatal care on one-year mortality. Below 1,500 grams, newborns receive the "very low birth weight" (VLBW) designation, which triggers discontinuously more intensive neonatal care protocols than newborns born above the threshold. Newborns just below 1,500 g are nearly biologically identical to those just above, yet receive meaningfully more care. The authors found that one-year mortality fell by approximately 1 percentage

point at the threshold (against a baseline of ~5.5% just above 1,500 g), suggesting that the intensified care protocol reduces mortality in this population.

What to read: Bor J et al. Regression discontinuity designs in epidemiology: causal inference without randomized trials. *Epidemiol*. 2014;25(5):729–737.

Non-randomized (observational) studies take another approach to attempt to make the exchangeability assumption believable. In these studies, researchers are usually forced to rely on modeling adjustments, rather than an actual physical mechanism, to approximate exchangeability. After adjusting for a sufficient set of measured confounders using regression, the hope is that the exposed and unexposed groups are comparable with respect to all other determinants of the outcome (i.e., the proportion of doomed and immune response types is the same in each group). Regression methods such as Cox proportional hazards models are often used to adjust for confounders and estimate the relationship between an exposure and outcome of interest. For example, Pan et al. (2012) used Cox regression, adjusting for age, BMI, smoking, alcohol, physical activity, diet quality, and family history of chronic disease, to estimate the association between daily red meat consumption and mortality in two large prospective US cohorts (the Nurses' Health Study, with N = 83, 644; and Health Professionals Follow-up Study, with N = 37, 698). They found that each additional daily serving of unprocessed red meat was associated with a 13% higher rate of total mortality (HR: 1.13; 95% CI: 1.07, 1.20), providing evidence that red meat consumption is associated with premature death.

All these methodologies for data analysis share an underlying goal: to attempt to ensure that the treated/exposed and untreated/unexposed groups are exchangeable, meaning that the proportion of doomed and immune response types is the same in each group (i.e., trying to make $q_1 = p_1$). When exchangeability holds, and in the presence of important additional assumptions,[4] the observed difference in outcomes between groups can be attributed to the treatment rather than to pre-existing differences between groups. The methodologies differ only in the mechanisms they rely on to invoke exchangeabilty in a credible way. Thus, when evaluating any of these methods, there is always a core question about whether you believe that the treated and untreated groups are truly comparable with respect to all determinants of the outcome, and thus whether exchangeability actually does hold? This is true whether one uses regression, interrupted time series, Mendelian randomization, or a range of other methods. Ideas that stem from randomization lie at the core of all of these methods.

[4] What are these? They are context specific, and discussed in the section on identification.

3.1 Randomized Trials Give More Than Just Exchangeability

However, randomization in an actual randomized trial gives us much more than exchangeability. The act of physically assigning individuals to a treatment strategy, as opposed to just observing individuals with particular treatment (or exposure) levels, affords researchers the opportunity to control a number of other issues or threats to validity beyond a lack of exchangeability. This is the reason that ideal randomized trials are often considered the "gold standard" for evaluating clinical efficacy and safety (or, more generally, for causal inference).

What to read:
Naimi et al Forthcoming Invited Commentary: Contextualizing Target Trial Emulation in Nutritional Epidemiology. *Am J Epidemiol*
Aronow et al 2025 Nonparametric identification is not enough, but randomized controlled trials are. *Obs Stud*;11(1):3-16

4 Randomized Controlled Trials and Target Trial Emulation

So far we have used randomization narrowly, as a device for confounding control, or a tool that allows us to equate potential response types in a given setting. But we've noted that a randomized trial does more than just give us exchangeability. What are those things, exactly?

We now turn to the *design* of a randomized trial: the protocol elements a trialist specifies before enrolling begins. These elements matter, because they supply the scaffolding for causal inference from data, and beyond randomized trials, can be used in the context of observational data through the idea of a **target trial**. When a randomized trial is unavailable or infeasible, we can still pose a causal question by describing the hypothetical trial we *would* run to answer it, and then using observational data to **emulate** this trial . The target trial is usually *hypothetical*, in the sense that it is the trial we would have run had it been feasible. However, a target trial can also be an actual, specific trial, as when an emulation is built to reproduce a completed randomized trial in order to benchmark the observational approach against a known answer.

What to read: Hernán MA, Robins JM. Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available. *Am J Epidemiol*. 2016;183(8):758-764.

The target trial emulation framework thus has two steps: (1) **specification** of the target-trial protocol, and (2) **emulation**, in which each protocol component is mapped onto the observed data (Hernán et al., 2026).

These protocol components organize the rest of this section: eligibility criteria, treatment strategies, the assignment (randomization) procedure, time zero and follow-up, outcomes, the causal contrast, and the identifying assumptions (Cashin et al., 2025). In a randomized trial these are usually given by design. In a trial emulation none are given, but knowing what protocol components are important and why they are important prepares us to identify

potential design errors, such as a misaligned time zero, eligibility assessed at the wrong moment, ill-defined treatment strategies, or other problems, that are the source of immortal-time and selection bias in observational analyses.[5] For each component below we note what it means in a trial and, in brief, what emulating it requires. Note that these protocol components have recently been codified in the TARGET Statement, which is a framework/checklist that allows us to be meticulous with our considerations.

[5] *Immortal time* is follow-up during which, by construction of the analysis, the event could not have occurred — for instance, when everyone analyzed had to survive to some later moment to be classified. Week 2 makes this concrete with the NHEFS data.

What to read: Cashin AG, Hansford HJ, Hernán MA, et al. Transparent Reporting of Observational Studies Emulating a Target Trial: The TARGET Statement. *JAMA*. 2025;334(12):1084-1093.

4.1 Eligibility Criteria

RCTs begin by delineating the eligible population.

Eligibility criteria are inclusion criteria (e.g., age range, diagnosis, disease severity) and exclusion criteria (e.g., contraindications, comorbidities, prior treatment) that shapes who gets into and who is left out of the trial. These criteria are usually considered in light of a tension between the "explanatory" versus "pragmatic" ideal in randomized trials (Schwartz and Lellouch, 1967).

What to read: Schwartz and Lellouch Explanatory and pragmatic attitudes in therapeutical trials. *J Chron Dis* 1967;20(8):637-48

Explanatory trials are meant to evaluate physiologic effects of treatment under "equalized" or (statistically) "optimal" conditions. In such a trial, the goal would be to reduce variability in the sample that might dilute physiologic effects. This leads to homogeneous RCT samples, and thus, a (sometimes severe) lack of generalizability of study results, an oft cited limitation of randomized trials (Cole and Stuart, 2010).

Pragmatic trials, on the other hand, are conducted to evaluate physiologic effects **in context**; that is, treatment effects in more practical, real-world settings. Rather than restricting enrollment to homogeneous, low-risk patients with the goal of reducing variability, pragmatic trials aim to recruit participants that reflect the population for which the treatment might be deployed at scale (Ford and Norrie, 2016). This is not meant to suggest that pragmatic trials are generalizable. Highly selective participation in research, including randomized trials, is a common problem.[6] However, as an ideal, pragmatic trials are aimed at understanding how the treatment or intervention *would work in practice*.

What to read: Ford I and Norrie J 2016 Pragmatic Trials. N Engl J Med 2016;375:454-63

[6] For example, a 2022 analysis from the National Cancer Institute found that only 9% of Americans reported having ever been invited to participate in a clinical trail, and of those, only 47% reported participating (National Cancer Institute, 2022).

There are several practical tools available to help researchers define eligibility criteria for a given trial. The PRECIS-2 tool includes nine domains that are meant to help trialists determine whether their design decisions align the the intended purpose (pragmatic versus explanatory) of the trial (Loudon et al., 2015).[7]

[7] The website `www.precis-2.org` contains useful information on how to use the tool, including a pdf toolkit linked here.

In emulation. With target trial emulation based on some observational or administrative dataset, eligibility can't be enforced at enrollment. It might be possible to *reconstruct* from measured covariates and individuals reported in the data, keeping only person-records that would have met each criterion at their own time zero. In the context of target trial emulation, the goal is to use eligibility criteria to identify which individuals in your data should be included in the analytic dataset, and to identify any individuals not in the source data that would have been included in the target trial. Importantly, this eligibility should be judged with respect to the study time zero, which is not always possible. However, conditioning on anything that happens after time zero (for instance, requiring that a patient survived long enough to initiate treatment) can introduce selection and/or immortal-time biases, which the TTE framework seeks to prevent.

4.2 Treatment Strategies, or the Trial Protocol / Objective

Treatment strategies define the study protocol, and are directly aligned with the research objectives. They determine what and how a particular treatment is allocated in the trial, and thus shape the question that is being answered by the trial. For example, in the Effects of Aspirin on Gestation and Reproduction (EAGeR) Trial, women were assigned to either take 81mg of Aspirin daily over the course of follow-up, or a matching placebo tablet. Participants were provided with a pill pack for the periods between clinical visits, roughly every two weeks during the "active follow up" period, and roughly every four weeks during the "passive follow up" phase. Participants were instructed to take the study medication daily throughout six menstrual cycles. If they became pregnant, they were instructed to take study medications until week 36 of pregnancy. Placebo tablets were manufactured to match on size, color, taste, and weight (Schisterman et al., 2013). Women in both treatment arms also received daily 400-mcg folic acid (generic). This EAGeR Trial treatment strategy is an example of a fixed strategy.

A dynamic strategy, often employed in the context of sequential multiple assignment randomized trial (SMARTs), specifies an assignment rule that adapts to the participant's evolving state. For example, Kasari et al. (2014) studied the effect of a blended developmental/behavioral intervention with or without the augmentation of a speech generating device for six months

of follow-up among minimally verbal children with autism 5-8 years old. The intervention consisted of two stages. First, children were randomized to either a behavioral language intervention (BLI) or BLI + augmentative and alternative communication approach (AAC). After three months of follow-up, children in each treatment arm were then categorized as responders or slow responders. Slow responders in the initial BLI group were then re-randomized to either an augmented BLI arm, or to an augmented BLI + AAC arm, and follow up was continued for an additional three months. The outcome of this study was based on a "Natural Language Sample" to measure a total number of spontaneous communicative utterances. Figure 4.2, taken directly from Lu et al. (2016), demonstrates how this dynamic treatment strategy was deployed.

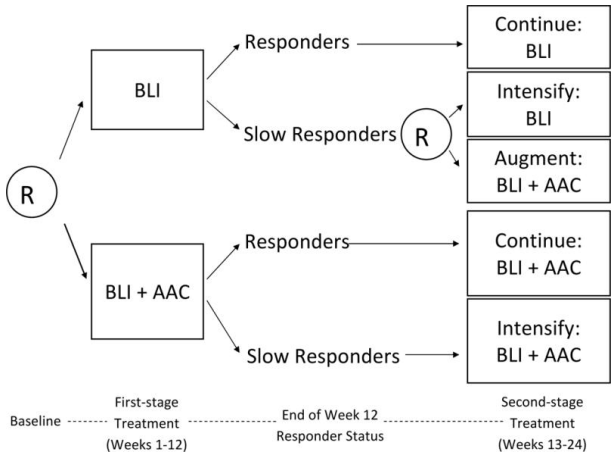

Figure 2: Design of the sequential multiple assignment randomized trial (SMART) conducted by Kasari et al. (2014) among minimally verbal children with autism. At baseline, children were randomized (first R) to a blended behavioral language intervention (BLI) alone or to BLI plus an augmentative and alternative communication (AAC) device. At the end of week 12 (first-stage treatment), children in each arm were classified as responders or slow responders. Responders continued their assigned first-stage treatment; slow responders to BLI alone were re-randomized (second R) to intensified BLI or to BLI augmented with the AAC device; and slow responders to BLI + AAC received intensified BLI + AAC. Second-stage treatment spanned weeks 13 to 24. Figure reproduced from Lu et al. (2016).

The treatment strategy must not only clearly specify the treatment, but also the comparator. Comparators include placebo comparators, active comparators.

- **Placebo**: In an RCT, placebo is commonly used, and, importantly, the placebo should be assigned in a manner that is as indistinguishable as possible as the actual treatment. Hence, in the EAGeR Trial, the trialists made it a point to note that "Placebo tablets were manufactured to match on size, colour, taste, and weight." (Schisterman et al., 2013, p600).
- **Active comparator**: Sometimes an active comparator is deployed. Active comparator designs can assess whether a new treatment is better than what is already in use. Consider tofacitinib for ankylosing spondylitis (AS). This new drug was known to result in short-term dose-dependent blood

changes and possessed a unique side-effect profile (very much unlike placebo). However, there is a separate class of drugs that were used in this clinical setting to treat AS, referred to as anti-Tumor Necrosis Factor (TNF) drugs. These drugs possess a similar side effect profile to JAK inhibitors, and should have been used as an active comparator in these trials.

In emulation. Observational data rarely contain explicit information on a "treatment strategy" of interest. Rather, researchers will rely on various devices that were used to construct the source data (e.g., whether a *pharmaceutical claim* was submitted to an insurance company or not) to define what constitutes being "on" versus "off" treatment (or to define the treatment dose), as well as to attempt to accurately classify the person-time under these treatment strategies or doses (i.e., "exposed" or "unexposed").

This leads to several real and potential complications in constructing an analytic variable representing the treatment of interest, and target trial emulation is meant to clarify what these are, and how a particular data source interacts with these complications (i.e., is this a problem with *our* data source). Some questions that should be considered when using some source observational data to construct an analytic variable representing the treatment strategy include:

1) Can the initiation of this treatment strategy, and the timing of the exposure measurement, align with time zero in my study?
2) How are the "exposed" and "unexposed" states, or the doses assigned, measured and defined?
3) For a given individual in my analytic data, can I be sure that all their person-time is accurately classified as exposed or unexposed?

4.3 Randomization and Assignment

The randomization mechanism in an RCT must also be clearly defined, and is the actual process by which participants are allocated to treatment arms.

Simple randomization is equivalent to a coin flip: each participant has a fixed probability (typically, but not necessarily, 0.5) of receiving treatment. In practice, this approach would proceed by creating a single randomization list before the trial is begun, and then each consecutive participant that's enrolled would receive the next treatment assigned by the list. However, in practice,

this has two important problems: First, simple randomization does not rule out repetitive sequences (for instance, several heads in a row), which can create important chance imbalances, especially in smaller trials. Second, and relatedly, it is possible that simple randomization fails to create balance across a variable that is an important predictor of the outcome, which can lead to "chance confounding" in the trial.

More principled and structured approaches are common, including:

Block randomization ensures treatment arms remain roughly balanced in size throughout the trial by randomizing within fixed-size blocks. In practice, a block size is chosen (e.g., 4 or 6), and all possible balanced sequences within that block are enumerated. For example, for block size of four with two arms (T, C), there are $\frac{4!}{(4/2)!(4/2)!} = 6$ sequences that can be used to allocate the four participants to treatment and control: (TTCC, TCTC, TCCT, CTTC, CTCT, CCTT).[8] One sequence is selected at random for each block, and patients are assigned to treatment in the selected order. This guarantees that after every completed block, the treatment arms are exactly balanced. Block size is often varied randomly (e.g., alternating blocks of 2, 4, and 6) to prevent investigators from predicting upcoming assignments once they know the block size.

[8] Note that the number of ways to arrange the blocks is a permutation problem, which can be calculated as $\frac{b!}{[(b/t)!]^t}$, where b is the block size, and t is the number of treatment arms. Hence, this design is sometimes referred to as a permuted block randomized trial design.

Stratified randomization randomizes separately within strata defined by key prognostic variables, or variables that can predict the outcome. In practice, the trialist first identifies one or more stratifying variables (e.g., site, sex, disease severity), then creates a separate randomization list (typically using a block structure) for each combination of strata. For example, with two sites and two severity levels, four independent block-randomized sequences are generated. Each participant is assigned to the appropriate stratum on enrollment and allocated to treatment using the sequence from that stratum's block. The key limitation is that the number of strata grows multiplicatively, so stratifying on more than two or three variables quickly yields strata that can be too sparse to maintain balance.

Adaptive randomization randomizes according to probabilities based on accumulating results. The most common implementation is *response-adaptive randomization* (e.g., the play-the-winner rule): after each observed outcome, the probability of assigning the next participant to the winning arm is increased. A Bayesian version maintains a posterior distribution over treatment effects and updates allocation probabilities at each interim. Adaptive randomization

raises practical concerns — it complicates inference, can introduce time trends as the allocation ratio shifts, and is most appropriate when ethical pressure to minimize exposure to the inferior arm is strong (e.g., pediatric or rare-disease trials).

In emulation. Observational data carry no randomization mechanism at all: no one *assigns* treatment, so we instead *classify* each eligible person into the strategy their data show they were following at time zero. The balance that results from randomization in expectation must therefore be assumed rather than engineered. For a time-fixed exposure measured at baseline, this typically occurs by adjusting for a sufficient set of pre-baseline confounders, to enable us to assume conditional exchangeability. The block-, stratified-, and adaptive-randomization machinery above has no counterpart in an emulation: its work is done instead by confounder measurement and adjustment, and by the other identifying assumptions noted below.

4.4 Blinding

Blinding conceals treatment assignment from participants, clinicians, and/or outcome assessors to prevent knowledge of assignment from influencing behavior or outcome ascertainment:

- **Single blinding**: participants are unaware of their assignment.
- **Double blinding**: both participants and treating clinicians are unaware.
- **Triple blinding**: participants, clinicians, and outcome assessors are all unaware.

Consider a phase III randomized double-blind placebo controlled trial of the effect of tofacitinib on ankylosing spondylitis (AS) among patients who have inadequately responded to or are intolerant of standard first line treatments (e.g., NSAIDs) (Deodhar et al., 2021). Ankylosing spondylitis is an arthritic condition that results in inflammation and fusing of the spinal column, leading to potentially severe pain and immobility. As a therapeutic agent, tofacitinib is an inhibitor of the Janus Kinase (JAK) pathway system, implicated in several autoimmune disorders closely related to AS (e.g., rheumatoid and psoriatic arthritis). Patients in the trial were randomized 1:1 to receive 5mg tofacitinib or placebo twice daily for 16 weeks of follow-up, and the primary study endpoint was based on a *self reported change score*, referred to as the Assessment of

SpondyloArthritis international Society ≥ 20% improvement (ASAS20) score. Both patients and clinicians were blinded to treatment assignment, and the study employed clinical assessors in an attempt to keep clinician investigators blinded.

However, tofacitinib is known to result in short-term dose-dependent changes in lipid concentrations, liver enzymes, creatine kinase, blood counts, as well as a specific side-effect profile. Results of tests for these markers, or the unique side-effect profile, could have led to functional unblinding of the study participants. It is possible that functionally unblinded participants in the treatment group reported better outcomes due to an expectancy effect (Huneke et al., 2025). Such expectancy effects, though not technically biases (Mansournia et al., 2017; Gabriel et al., 2025), would threaten the validity of a study on a drug like tofacitinib. The potential for such an expectancy effect could have been mitigated via an active comparator design (Lund et al., 2015). Another more recent clinical question that trials have a hard time answering because of unblinding is the effect of psychedelics such as psilocybin on mental health outcomes such as depression and anxiety, whether compared to inert placebo, to an active placebo such as methylphenidate (chosen in early psilocybin studies precisely to make blinding more credible), or to an active comparator such as escitalopram. Being assigned to psilocybin can lead to functional unblinding, leading to beliefs about the drug's effect and expected outcomes, which is opening a new realm of research in causal inference (Loewinger et al., 2025).

Blinding matters most when outcomes are subjective or self-reported (pain, quality of life, self-reported change in condition score) or when clinicians might differentially apply co-interventions.

In emulation. There is no analogue to blinding when we use observational data to construct an analytic dataset that seeks to emulate a target trial. However, it may be possible to construct treatment strategies that can offset the fact that some people know what they are exposed to in a target trial emulation. For instance, one can use observational data to construct treatment and referent categories that align with an active-comparator design. One can and should attempt to identify what might equate to "new-users" in an observational data source, to mitigate the complications that result from analyzing data from prevalent users (Yoshida et al., 2015).

4.5 Time Zero and Follow-up

Time zero is the *moment we start recording events in the study (the moment follow-up begins),* and one of the key benefits of randomization is that there is a relatively clear anchor that we can use to set time zero. Most commonly, the date in which the subject was randomized is taken as time zero. However, this is not always the case.

Setting time zero also means committing to a **timescale**, the clock along which follow-up is measured. The start time should correspond to some well-defined event such as an age of interest (age as timescale), a date of interest (calendar date as timescale), or the timing of some important study marker (e.g., date of randomization to treatment versus placebo). Consider the following examples from the literature, each with a different study timescale:

1) Naimi et al. (2021) use data from a randomized trial to estimate the adherence adjusted per protocol effect of daily low-dose aspirin on pregnancy outcomes in ~1,200 women. In this study, the timescale was **weeks since randomization**, and ranged from 0 to 60 weeks.

2) Getahun et al. (2005) examined stillbirth, small for gestational age, and infant mortality occurrence by the racial classification of both parents (e.g., white-white, white-black, black-white, black-black) in roughly 20 million pregnancies in the United States. In this study, the timescale was **gestational age**, starting at the 20th week of gestation.

3) Huang et al. (2018) looked at the relation between different post-operative management strategies, including the use of dexamethasone versus flurbiprofen axetil on survival in 588 patients undergoing surgical lung resection for non-small-cell lung cancer. In this study, the timescale was **time since surgical resection**.

4) Schwarzinger et al. (2018) looked at the relation between alcohol use and dementia risk in nearly 31 million individuals in France between 2008 and 2013. In this analysis, the timescale was **age**. The origin of the age timescale is birth (age zero), but no one in the study was followed from birth: each individual instead *entered* follow-up at whatever age they had reached when the study period opened. This gap between a timescale's origin and the moment a person actually comes under observation is called

delayed (or late) entry, and it must be accounted for in the analysis; we return to it when we discuss left truncation below.

5) Sabia et al. (2019) looked at the association between cardiovascular health at age 50 and the risk of subsequent dementia in ~8,000 individuals enrolled in the Whitehall II study. In this analysis, the timescale was **calendar date**: person-time was indexed by the calendar itself, with each participant entering follow-up on the date of their clinical examination at age 50 (a date that differs from person to person) and followed forward until dementia, death, or the end of follow-up. Aligning person-time this way means that individuals compared at the same analysis time share the same period (for example, the same diagnostic practices), rather than the same age or the same duration of follow-up.

In landmark analysis trials, a specific time-point "landmark" (e.g., 6 months after starting treatment) is selected, and only patients who survived or remained event-free up to that landmark are included in the analysis (Morgan, 2019). In these cases, time zero is effectively shifted from the randomization date to the landmark date, with the intent of ensuring the "event" being measured was not actually a "pre-randomization" event. In the past (less commonly now), landmark analyses were used to deal with time-dependent covariates, but better methods are available. These include time-varying methods such as Robins' G methods (Naimi et al., 2017; Robins and Hernán, 2009) or clone-censor-weighting approaches (Cain et al., 2010).

In trials with run-in periods, patients may be randomized, but then undergo a "run-in" period (e.g., a placebo run-in to test adherence) before starting the actual intervention. If the study focuses on the effect of the intervention itself, the time zero for analysis may be the day they start the active drug, not the day they were randomized into the run-in phase.

In sequentially nested trials, there are several time zeros that define the initiation of the trial "nested" within the trial. For example, Figure 4.2 is an example, where the second trial is nested within an arm of the first trial.

Mishandling of time zero is one of the most common sources of bias in observational studies (immortal time bias, prevalent user bias), which, as we'll discuss, is why explicitly anchoring an emulated trial to a well-defined time zero is so important. In randomized trials, this is easier to do.

In emulation. In emulation, time zero is something usually *imposed* or selected by researchers. There is a cardinal "alignment" rule that may or may not be possible to enforce: the moment that research participants (i.e., sample observations) are classified into a strategy should coincide with the moment their follow-up begins. This may not always be possible (say follow-up starts before the treatment strategy begins), and problems such as left truncation or immortal time bias arise as potential problems. Anchoring eligibility, assignment, and start of follow-up to a single, explicitly defined time zero is the framework's principal safeguard against this and the related prevalent-user bias. The key is to think about what the data would "look like" if you were able to actually run the trial, and randomize participants to the dose or group you are interested in studying. As you start to identify where your data cannot look like what the trial data would look like, you begin to identify some of the potential threats to the validity of your study.

4.6 Outcomes

Outcomes are usually grouped into a *primary* outcome (the outcome the trial is powered to detect and that anchors the main analysis) and *secondary* outcomes. Specifying an outcome means stating not only *what* is measured but *how* and *when*: the instrument or definition, the ascertainment schedule, and the window over which events are counted. In the tofacitinib trial, for example, the primary endpoint was a self-reported change score (ASAS20) assessed at a fixed 16-week visit, which is a subjective, time-anchored measurement. This is is precisely why blinding mattered much more in this trial that it might in others.

Specifying an outcome also requires being explicit about the other things that can happen before or during follow-up, that can bias estimates of interest. Two key considerations include **censoring** and **truncation**, which are processes that can prevent us from seeing an event that occurred, and whether a person that should be in our study is actually in our study.

A third key consideration is whether **competing events** are present. This issue is about whether the outcome of interest can happen at all, and, if not, what we should do about analyzing our data.

In a trial setting, all three are design questions in that the protocol has to anticipate them before the first participant is enrolled.

Censoring and Truncation

Censoring and truncation, can both be classified as left, right, or interval censoring/truncation. But, most commonly, in epidemiologic studies we encounter left and right censoring (each of which has different implications for our analysis), and left truncation.

Censoring occurs where we have information on the individual, they are in our study. However, we either don't see whether the outcome(s) occurred (right censoring), OR we don't know precisely when the outcome occurred (but we know whether they experienced the outcome or not; left and interval censoring).

Truncation occurs when an individual should have been in our study, but because of how we define the start or end of follow up, we never have the opportunity to see them in our data. Truncation can also be classified as right, left, or interval, with the most common being left truncation.

Both can lead to a form of selection bias because, if we ignore censoring or truncation, we end up analyzing data on a selected group of individuals that may yield misleading results.

The processes behind censoring and truncation are easiest to see on a timeline. The figure below shows six observations in a hypothetical study, each drawn along the study timescale. The shaded box is the window over which the study collects data, solid segments are person-time we actually observe, and dashed segments are hidden person-time. Observations with IDs 1, 2, and 3 depict right, left, and interval censoring, respectively. Observations with IDs 4, 5, and 6) depict right, left, and interval truncation, respectively.

Right Censoring (ID = 1) occurs when an individual is enrolled in the study, but we don't know whether the individual has had an event of interest or not. This type of censoring often occurs because either an enrolled individual leaves the study (withdrawal, or loss to follow-up), or the study ends for everyone at a common cutoff, unrelated to any individual's prognosis (administrative censoring). This distinction is an important one, which will come up several times in the class. Right censoring is often said to be the most common type of censoring. Generally, when we use the word "censoring" in this class, we are referring to right censoring.

Left Censoring (ID = 2) occurs when an individual is enrolled in the study, and we know has experienced an event of interest (and we know which event

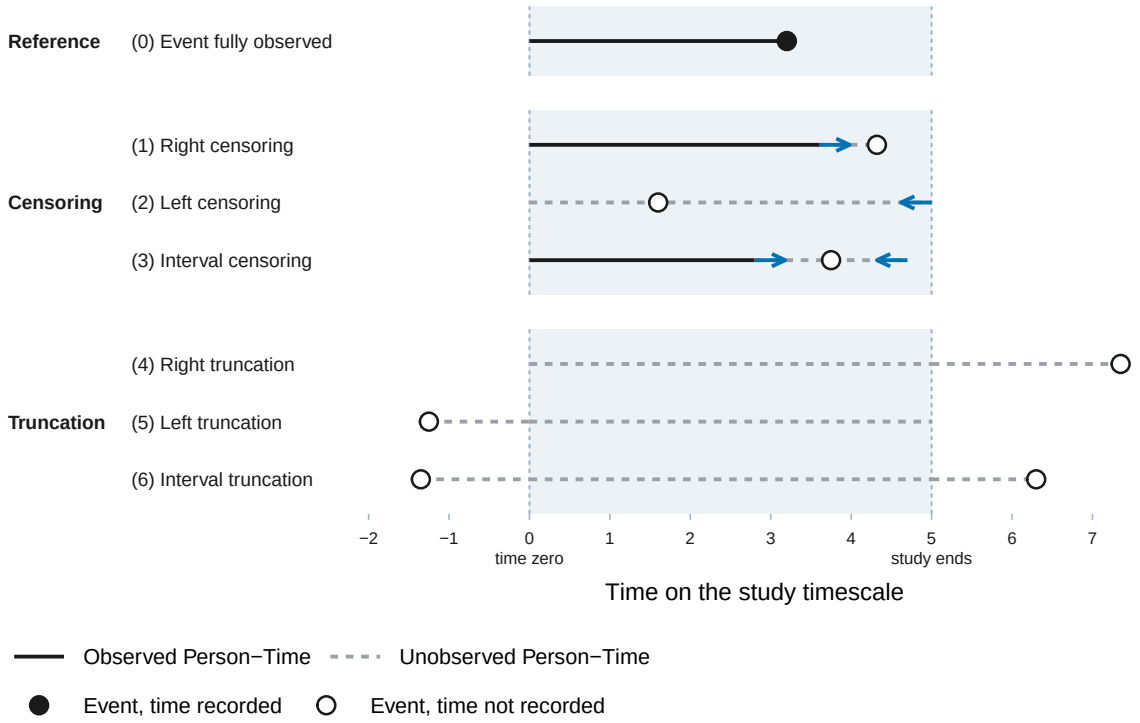

Figure 3: Six observations in a hypothetical study depicting censoring and truncation (left, right, and interval for both).

it is), but we have no information on *when* the event occurred. I believe this to be the most common type of censoring, due to the fact that most often, we collect data on whether an event occurred or not during the course of our study, and not on the exact timing of events. Thus, outcomes in a typical cohort study that do not have information on the timing of events are left censored. Practically, this means that we can construct risk estimates, and functions of risk estimates (e.g., odds), at the end of follow up. We cannot, however, construct measures of occurrence that depend on time, such as cumulative risk functions, hazards, or other measures that require information on the exact timing of the event.

Interval Censoring (ID = 3) occurs when an individual is enrolled in the study, and we know has experienced an event of interest (and we know which event it is), but we only know that the event occurred in a bounded *interval*, with the bounds occurring after the study start date and before the study end date. As with left censoring, this means that we can construct risk estimates, and functions of risk estimates (e.g., odds), at the end of follow up. However, depending on how wide the intervals are, we may or may not be able to construct measures of occurrence that depend on time.

Right Truncation (ID = 4) occurs when an individual is NOT enrolled in the study because the event happened after a particular date. One example is in Medley et al. (1987), who studied time from exposure to HIV contaminated blood or blood products and the development of AIDS. Data were collected retrospectively from individuals with confirmed AIDS diagnosis. The number of individuals who were exposed to HIV contaminated blood or blood products that had not yet developed AIDS was not known. In this study, only those individuals who developed AIDS by the time the study was enrolling could be identified for inclusion, which resulted in right truncated data.

Left Truncation (ID = 5) occurs when an individual is NOT enrolled in the study because the event happened before a particular date. This type of truncation is common in studies of spontaneous abortion. For example, Waller et al. (1998) examined the relation between prenatal exposure to trihalomethanes in drinking water (a by product of chlorination) and spontaneous abortion. Women were recruited from prenatal care clinics. However, spontaneous abortion tends to be more common early in pregnancy (and can often be confused with normal menstruation). Thus, it is likely that many sponta-

neous abortions were missed because they occurred before enrollment began, resulting in left truncated data.

Interval (Double) Truncation (ID = 6) occurs when an individual is NOT enrolled in the study because the event happened between two dates. Interval truncation occurs in studies of, for example, autopsy confirmed neurodegenerative diseases (ND). On the one hand, diagnosing ND is difficult, and studies tend to focus on the occurrence of disease in older populations. Thus, individuals who experience ND early tend not to be included in these studies. On the other hand, because autopsy confirmation is required for inclusion in the study, individuals who survive past the study end date are also not included. This example, as well as methods to address interval truncation, are discussed in Rennert (2018).

It's important to connect the idea of censoring and truncation to our ability to validly interpret regression model parameters as risk differences, risk ratios, or odds ratios.[9] Clearly, censoring and truncation matter because they determine whose outcome is observed or who is in cohort. Without carefully considering how to handle censored or truncated data, we can obtain biased results. Unfortunately, it's not always possible to appropriately handle censoring and truncation. Methods do exist, and we will cover some of them when we turn to survival analysis in Week 12 (time permitting), but it's not always possible to use these methods in the context of a given study. Thus, understanding that these issues might be present in a dataset, but inappropriately dealt with, becomes one of the central limitations that should be discussed in a scientific project.

[9] Validity here depends on more than just the presence or absence of censoring and truncation. But appropriate handling of censoring and truncation (i.e., necessary, but not sufficient).

Competing Events / Risks

A competing event is one whose occurrence prevents the outcome of interest from ever occurring. This could be (is most commonly) death, or death from another cause when the outcome itself is death from a specific cause. Competing events are not censoring mechanisms, and the distinction between these is important. A censored individual is, as far as we know, still at risk for the event under study, we've simply stopped observing them. An individual who experiences a competing event can never experience the event of interest. Which events count as competing, and therefore which risk the trial is after, is a decision that belongs in the protocol alongside the outcome definition itself.

The presence of competing events in a study leads to an important distinc-

tion in the types of risk we can estimate. In effect, the concept of risk can be re-categorized depending on how we want to handle a competing event.

The estimand changes.[10]

The first estimand is the **cause-specific risk**: the risk of the outcome of interest we would see if the competing event could be switched off, keeping everyone at risk for the event of interest. There are several ways to estimate this risk. For instance, we could use the Kaplan–Meier estimator, treating the competing (other-cause) deaths as though they were censored[11] (Lau et al., 2009).

The second estimand is the **sub-distribution risk**, otherwise known as the "cumulative incidence function."[12] The sub-distribution risk for an outcome of interest is the probability of that outcome *in the real world, where competing risk is allowed to happen* (Cole et al., 2015; Lau et al., 2009). Someone who experiences a competing risk at (say) year 5 can never experience the event of interest later on in the study. There are several ways too estimate sub-distribution risks, with the most common approach being the Aalen–Johansen estimator (which generalizes KM to multiple event types).

We cover these issues in more depth later, but for now (from a trial design perspective), note that it is important to articulate which risk we are interested in quantifying as a part of the protocol. Therefore, in building a trial, one should have a sense of the competing risks that a study participant can experience, and whether our scientific interest lies in cause-specific or sub-distribution estimands.

In emulation. With observational data an outcome is often not measured according to a pre-specified protocol, but reconstructed from records (such as diagnosis codes, laboratory values, registry linkages, or other "convenience" metrics) so that the outcome is often defined algorithmically. When doing this, it is essential to attempt to capture all the relevant forces leading to whether a person get's classified as an event of interest or not. In a trial emulation, explicit consideration should be made with regards to censoring and truncation mechanisms, and whether a competing event is possible. If possible, efforts should be made to measure the competing events, and then incorporate them into the analysis appropriately (via tools cause-specific or sub-distribution estimation).

[10] An *estimand* is the precisely defined quantity we are trying to estimate — here, a particular risk function. Week 2 defines estimands formally and distinguishes them from the *estimators* we use to compute them from data.

[11] You may have met this in EPI 545 as the "conditional" risk The KM machinery is identical to the single-event case; only the coding of the competing event as censored is new.

[12] This terminology, dating, as far as I am aware, to the original edition of Kalbfleisch and Prentice (2011), is unfortunate and confusing, since the terms "cumulative risk" and "cumulative incidence" are often used interchangeably, and to convey a number of different notions of risk. However, in the context of competing events / risks, it's useful to know that reference to "cumulative incidence" is often meant to convey "cumilative sub-distribution risk" functions.

4.7 Causal Contrasts: Intention-to-Treat and Per-Protocol

The main benefit of randomization, as noted, is that the randomization indicator (e.g., whether the individual was assigned to treatment or placebo) indexes groups that are exchangeable in expectation. However, the randomization indicator does not always perfectly align with the treatment itself. Instead, it aligns with an intention to treat the participant with the treatment, or the placebo (represented by the green node in Figure 1). In many trials, however, there may be a possibility that participants fail to follow what they were randomized to, and there are several unique circumstances in which this might play out.

For instance, in the EAGeR trial, participants were randomized to daily low dose aspirin or daily placebo. However, on any given day, a participant randomized to the aspirin arm may or may not have take the assigned aspirin pill. Similarly, on any given day, a participant randomized to the placebo group may or may not have taken the assigned placebo pill.

If we assume that there was only a single opportunity (time-point) to take, or not take, the aspirin or placebo pills, we can construct variables to represent a number of scenarios. For instance, we let X denote whether a participant was randomized to treatment ($X = 1$) or placebo ($X = 0$), and we let A denote whether the participant took aspirin ($A = 1$), or took placebo ($A = 0$). The potential relationships between these variables is depicted in Figure 4.7.

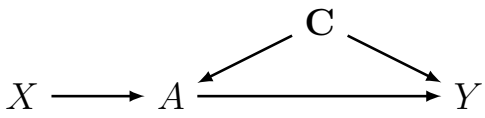

Figure 4: Causal diagram representing a randomized controlled trial design, where X denotes the randomization indicator, A denotes the treatment actually taken (aspirin, $A = 1$, versus placebo, $A = 0$), C denotes a confounder of the treatment taken and the outcome, and Y the outcome of interest.

In classical randomized trial analyses, researchers typically present one of three results. The intention-to-treat effect estimates might be obtained by comparing the mean outcome among those with $X = 1$ compared to those with $X = 0$. As can be gleaned from the DAG, this effect is identified due to the randomization process.

However, in the past researchers typically presented naive “as treated” and “per protocol” effects as:

$$E(Y \mid A = 1) - E(Y \mid A = 0)$$

and

$$E(Y \mid X = 1, A = 1) - E(Y \mid X = 0, A = 0)$$

with the hopes of better capturing the "physiological" effect of aspirin itself (rather than the effect of being asked to take aspirin, or the intention to treat). Unfortunately, as can be seen in the DAG, neither of these effects is identified unless one accounts for confounding the relation between A and Y.Therefore, even in randomized trials, the *naive* "as treated" and "per protocol" effects are unadjusted effects in potentially confounded observational studies.

More principled analyses will often adjust for necessary confounding variables, and present the effect of following a protocol of interest (i.e., the per protocol effect) (Rudolph et al., 2020).

In emulation. An emulated trial has no randomization indicator X, so the intention-to-treat contrast (the effect of being randomly assigned to a treatment) has no direct counterpart. Nevertheless, published trial emulations have routinely targeted what is often referred to as a "modified intention to treat" effect (e.g., Caniglia et al., 2020). It's important to recognize that this shares no relation to the intention-to-treat effect, since there is no assignment procedure in observational data. Identifying this "modified" intention-to-treat, or the accompanying per protocol effects that are also targeted in trial emulation, requires that the identification assumptions needed to estimate causal effects in observational data hold.

4.8 Identifying Assumptions

One of the bigger motivations for using target trial emulation framework in the first place is that it allows us to clarify exactly what conditions need to be in place for us to interpret observational results as though they were from a trial, to understand what we need to do in our observational data to make these conditions more likely in our observational setting, AND to identify where our observational data clearly or likely fail at meeting these conditions. This latter stipulation allows us to either (i) (preferably) implement formal sensitivity or robustness analyses which provide an indication of how important these

conditions/assumptions are to our analyses, or (ii) more clearly and impactfully outline the limitations of the work we are doing so that future work can improve upon it.

There are different sets of identification assumptions we can rely on to interpret observational study results as though they were generated from a randomized trial. The most common set invoked among the different sets include:

Conditional exchangeability. Within levels of a sufficient set of measured covariates, those following each strategy must have the same distribution of potential outcomes. This is often connected to the "no unmeasured confounding" assumption, but also implicates other forms of bias including selection bias and measurement error. For a per-protocol contrast in which strategies unfold over time, this strengthens to *sequential* exchangeability: it must hold at each point where treatment can change, conditional on the covariate history up to that point.

Positivity. Within every joint stratum that arises from all covariates adjusted for, each treatment/exposure must have a nonzero probability of being followed. This translates to there being sufficient exposed and unexposed individuals in each stratum. A covariate pattern under which one strategy is never observed leaves that stratum's contrast unidentified from the data alone.

Consistency (well-defined strategies). The potential outcome under a strategy must be the outcome we would in fact observe for someone who followed it. This presupposes that the strategy is specified precisely enough to be well defined, which is one reason why the "treatment strategies" domain is important in the context of target trial emulation.

The value of the target-trial emulation framework is that it clarifies *where* these assumptions are binding, and whether they are likely to hold in the observational data we are using to infer effects.

In an ideal randomized trial the intention-to-treat effect earns all three by design: randomization delivers unconditional exchangeability, the assignment mechanism guarantees positivity, and a well-specified protocol gives consistency. This is one reason why an RCT needs no covariate adjustment for its primary contrast. Note particularly here that it is not just exchangeability we obtain from randomization, but all three of these assumptions, and these, collectively, are what make an RCT the "gold standard" for causal inference.

An emulation inherits none of this automatically. The assumptions become claims about the data and the measured covariates that must be argued or interrogated, and reported, each on their own terms.

5 Conclusion

We began with randomization as a device for "extraction": a way of lifting an exposure-outcome relation out of the web of causes in which it is normally embedded. Following Greenland and Robins (1986), we saw what this buys us mathematically. Individuals can be classified into one of four response types (doomed, causative, preventive, immune), and a comparison of risks between the treated and untreated reduces to $(p_1 + p_2) - (q_1 + q_3)$. Without some assurance that the distribution of response types is the same in both group (i.e, that $p_1 = q_1$) a nonzero risk difference does not suggest that exposure caused anything at all. Exchangeability is the assurance we need, and physical randomization is what makes it credible, in expectation.

This framing let us place a range of designs on common ground. Mendelian randomization leans on the random segregation of alleles at meiosis; interrupted time series on a policy change whose timing is exogenous with respect to the outcome trend; regression discontinuity on a threshold that treats near-identical individuals differently; and conventional observational analysis on measured covariates and a regression model. These approaches differ in the mechanism they invoke, but not in the goal of equalizing the distribution of response types to achieve exchangeability. In every case, the question a reader should be asking is the same: do I believe the treated and untreated groups are comparable with respect to all determinants of the outcome?

We also saw that an actual randomized trial delivers more than exchangeability. The second half of these notes was devoted to what this consists of. A trialist has to write a protocol before enrolling anyone: who is eligible, what treatment strategies are being compared, how treatment assignment happens, who is blinded and how, when the follow-up clock starts and on what timescale, what outcome is measured and how censoring, truncation, and competing events will be handled, and which contrast (intention-to-treat or per protocol) the trial is after.

Each of these decisions does some part of the causal work, beyond what

randomization alone can do.

The target trial emulation framework takes these protocol components and turns them into a framework for observational analysis. It specifies the trial you would have run, then map each component onto the data you actually have (Hernán et al., 2026; Cashin et al., 2025). Its value is less that it makes observational data behave like trial data than that it makes the gaps explicit: Eligibility has to be reconstructed rather than enforced. Treatment strategies have to be inferred from claims, prescriptions, or diagnostic codes rather than administered. Assignment has no counterpart, but part of what it does is redistributed to confounder measurement and adjustment. Blinding has no counterpart either, though active-comparator and new-user designs can absorb some of what it was protecting against. Time zero must be *imposed*, and imposed such that eligibility assessment, strategy classification, and the start of follow-up all coincide. Outcomes are defined according to some rules applied to the data (like treatment strategies), so information on censoring, truncation, and competing events has to be made explicit and sought in the observational source (and one has to commit to the risk estimands that might result).

References

Douglas Almond, Joseph J. Doyle, Amanda E. Kowalski, and Heidi Williams. ESTIMATING MARGINAL RETURNS TO MEDICAL CARE: EVIDENCE FROM AT-RISK NEWBORNS. *The quarterly journal of economics*, 125(2):591–634, 2010.

Stephen Burgess, Simon G. Thompson, and CRP CHD Genetics Collaboration. Avoiding bias from weak instruments in Mendelian randomization studies. *International Journal of Epidemiology*, 40(3):755–764, 2011.

Lauren E. Cain, James M. Robins, Emilie Lanoy, Roger Logan, Dominique Costagliola, and Miguel A. Hernán. When to start treatment? A systematic approach to the comparison of dynamic regimes using observational data. *The International Journal of Biostatistics*, 6(2):Article 18, 2010. ISSN 1557-4679. DOI: 10.2202/1557-4679.1212.

Ellen C. Caniglia, L. Paloma Rojas-Saunero, Saima Hilal, Silvan Licher, Roger Logan, Bruno Stricker, M. Arfan Ikram, and Sonja A. Swanson. Emulating a target trial of statin use and risk of dementia using cohort data. *Neurology*, 95(10):e1322–e1332, September 2020. ISSN 0028-3878. DOI: 10.1212/WNL.0000000000010433.

Aidan G. Cashin, Harrison J. Hansford, Miguel A. Hernán, Sonja A. Swanson, Hopin Lee, Matthew D. Jones, Issa J. Dahabreh, Barbra A. Dickerman, Matthias Egger, Xabier Garcia-Albeniz, Robert M. Golub, Nazrul Islam, Sara Lodi, Margarita Moreno-Betancur, Sallie-Anne Pearson, Sebastian Schneeweiss, Melissa K. Sharp, Jonathan A. C. Sterne, Elizabeth A. Stuart, and James H. McAuley. Transparent Reporting of Observational Studies Emulating a Target Trial—The TARGET Statement. *JAMA*, 334(12):1084–1093, September 2025. ISSN 0098-7484. DOI: 10.1001/jama.2025.13350.

Stephen R. Cole and Elizabeth A. Stuart. Generalizing Evidence From Randomized Clinical Trials to Target Populations: The ACTG 320 Trial. *Am J Epidemiol*, 172(1):107–115, 2010.

Stephen R. Cole, Michael G. Hudgens, M. Alan Brookhart, and Daniel Westreich. Risk. *Am J Epidemiol*, 181(4):246–250, 02 2015.

George Davey Smith and Gibran Hemani. Mendelian randomization: Genetic anchors for causal inference in epidemiological studies. *Human Molecular Genetics*, 23(R1):R89–98, 2014.

Atul Deodhar, Paula Sliwinska-Stanczyk, Huji Xu, Xenofon Baraliakos, Lianne S. Gensler, Dona Fleishaker, Lisy Wang, Joseph Wu, Sujatha Menon, Cunshan Wang, Oluwaseyi Dina, Lara Fallon, Keith S. Kanik, and Désirée van der Heijde. Tofacitinib for the treatment of ankylosing spondylitis: A phase III, randomised, double-blind, placebo-controlled study. *Annals of the Rheumatic Diseases*, 80(8):1004–1013, August 2021. ISSN 1468-2060. DOI: 10.1136/annrheumdis-2020-219601.

Ian Ford and John Norrie. Pragmatic trials. *!!! no full title or abbreviation for line:*, 375(5):454–463, 2016.

Erin E. Gabriel, Alex Ocampo, and Arvid Sjölander. Elucidating some common biases in randomized controlled trials using directed acyclic graphs. *European Journal of Epidemiology*, September 2025. ISSN 1573-7284. DOI: 10.1007/s10654-025-01298-7.

Darios Getahun, Cande V. Ananth, Nandini Selvam, and Kitaw Demissie. Adverse perinatal outcomes among interracial couples in the united states. *Obstetrics & Gynecology*, 106(1), 2005.

Sander Greenland and JM Robins. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol*, 15(3):413–419, 1986.

Gibran Hemani, Jack Bowden, and George Davey Smith. Evaluating the potential role of pleiotropy in Mendelian randomization studies. *Human Molecular Genetics*, 27(R2):R195–R208, August 2018. ISSN 0964-6906. DOI: 10.1093/hmg/ddy163.

M. A. Hernán and JM Robins. *Causal Inference*. Chapman & Hall/CRC, Boca Raton, FL, 2020.

Miguel A. Hernán, Barbra A. Dickerman, Sonja A. Swanson, and Issa J. Dahabreh. Where Do Target Trials Come From? Specifying the Protocol of a Target Trial When Repurposing Data for Causal Inference. *Epidemiology*, 37(3):282–286, May 2026. ISSN 1531-5487. DOI: 10.1097/EDE.0000000000001951.

Wen-Wen Huang, Wen-Zhi Zhu, Dong-Liang Mu, Xin-Qiang Ji, Xiao-Lu Nie, Xue-Ying Li, Dong-Xin Wang, and Daqing Ma. Perioperative management may improve long-term survival in patients after lung cancer surgery: A retrospective cohort study. *Anesth Analg*, 126(5):1666–1674, 2018.

Nathan T. M. Huneke, Guilherme Fusetto Veronesi, Matthew Garner, David S. Baldwin, and Samuele Cortese. Expectancy Effects, Failure of Blinding Integrity, and Placebo Response in Trials of Treatments for Psychiatric Disorders: A Narrative Review. *JAMA Psychiatry*, 82(5):531–538, May 2025. ISSN 2168-622X. DOI: 10.1001/jamapsychiatry.2025.0085.

John D Kalbfleisch and Ross L Prentice. *The statistical analysis of failure time data*. John Wiley & Sons, Hoboken, NJ, 2011.

Connie Kasari, Ann Kaiser, Kelly Goods, Jennifer Nietfeld, Pamela Mathy, Rebecca Landa, Susan Murphy, and Daniel Almirall. Communication Interventions for Minimally Verbal Children With Autism: Sequential Multiple Assignment Randomized Trial. *Journal of the American Academy of Child and Adolescent Psychiatry*, 53(6):635–646, 2014.

Bryan Lau, Stephen R. Cole, and Stephen J. Gange. Competing risk regression models for epidemiologic data. *Am J Epidemiol*, 170(2):244–256, 2009.

Gabriel Loewinger, Mats J. Stensrud, Sandeep M. Nayak, David Yaden, and Alexander W. Levis. Causal Inference in Studies with Functional Unmasking: Psychedelics and Beyond, December 2025.

Kirsty Loudon, Shaun Treweek, Frank Sullivan, Peter Donnan, Kevin E. Thorpe, and Merrick Zwarenstein. The PRECIS-2 tool: Designing trials that are fit for purpose. *BMJ*, 350:h2147, 2015.

Xi Lu, Inbal Nahum-Shani, Connie Kasari, Kevin G. Lynch, David W. Oslin, William E. Pelham, Gregory Fabiano, and Daniel Almirall. Comparing Dynamic Treatment Regimes Using Repeated-Measures Outcomes: Modeling Considerations in SMART Studies. *Statistics in medicine*, 35(10):1595–1615, 2016.

Jennifer L. Lund, David B. Richardson, and Til Stürmer. The active comparator, new user study design in pharmacoepidemiology: Historical foundations and contemporary application. *Current epidemiology reports*, 2(4):221–228, 2015.

Mohammad Ali Mansournia, Julian P. T. Higgins, Jonathan A. C. Sterne, and Miguel A. Hernán. Biases in randomized trials: A conversation between trialists and epidemiologists. *Epidemiology (Cambridge, Mass.)*, 28(1):54–59, January 2017. ISSN 1044-3983. DOI: 10.1097/EDE.0000000000000564.

G. F. Medley, R. M. Anderson, D. R. Cox, and L. Billard. Incubation period of aids in patients infected via blood transfusion. *Nature*, 328(6132):719–721, 1987.

Charity J. Morgan. Landmark analysis: A primer. *Journal of Nuclear Cardiology*, 26(2):391–393, 2019.

Ashley I Naimi, Stephen R Cole, and Edward H Kennedy. An Introduction to G Methods. *Int J Epidemiol*, 46(2):756–762, 2017.

Ashley I. Naimi, Neil J. Perkins, Lindsey A. Sjaarda, Sunni L. Mumford, Robert W. Platt, Robert M. Silver, and Enrique F. Schisterman. The Effect of Preconception-Initiated Low-Dose Aspirin on Human Chorionic Gonadotropin-Detected Pregnancy, Pregnancy Loss, and Live Birth : Per Protocol Analysis of a Randomized Trial. *Annals of internal medicine*, 174(5): 595–601, May 2021. ISSN 1539-3704 0003-4819. DOI: 10.7326/M20-0469.

National Cancer Institute. Clinical Trial Participation Among US Adults. Health Information National Trends Survey Number 48, National Cancer Institute, 2022. URL https://hints.cancer.gov/docs/Briefs/HINTS_Brief_48.pdf.

Babak J. Orandi, M. Chandler McLeod, Paul A. MacLennan, William M. Lee, Robert J. Fontana, Constantine J. Karvellas, Brendan M. McGuire, Cora E. Lewis, Norah M. Terrault, Jayme E. Locke, and US Acute Liver Failure Study Group. Association of FDA Mandate Limiting Acetaminophen (Paracetamol) in Prescription Combination Opioid Products and Subsequent Hospitalizations and Acute Liver Failure. *JAMA*, 329(9):735–744, 2023.

An Pan, Qi Sun, Adam M. Bernstein, Matthias B. Schulze, JoAnn E. Manson, Meir J. Stampfer, Walter C. Willett, and Frank B. Hu. Red meat consumption and mortality: Results from 2 prospective cohort studies. *Archives of Internal Medicine*, 172(7):555–563, 2012.

Lior Rennert. *Statistical Methods For Truncated Survival Data*. PhD thesis, University of Pennsylvania, 2018.

James M Robins and Miguel Á Hernán. Estimation of the causal effects of time-varying exposures. In G Fitzmaurice, M. Davidian, G Verbeke, and G Molenberghs, editors, *Advances in Longitudinal Data Analysis*, pages 553–599. Chapman & Hall, Boca Raton, FL, 2009.

Jacqueline E. Rudolph, Ashley I. Naimi, Daniel J. Westreich, Edward H. Kennedy, and Enrique F. Schisterman. Defining and Identifying Per-protocol Effects in Randomized Trials. *Epidemiology (Cambridge, Mass.)*, 31(5):692–694, September 2020. ISSN 1531-5487 1044-3983. DOI: 10.1097/EDE.0000000000001234.

Séverine Sabia, Aurore Fayosse, Julien Dumurgier, Alexis Schnitzler, Jean-Philippe Empana, Klaus P Ebmeier, Aline Dugravot, Mika Kivimäki, and Archana Singh-Manoux. Association of ideal cardiovascular health at age 50 with incidence of dementia: 25 year follow-up of whitehall ii cohort study. *BMJ*, 366:l4414, 2019.

Eleanor Sanderson, M. Maria Glymour, Michael V. Holmes, Hyunseung Kang, Jean Morrison, Marcus R. Munafò, Tom Palmer, C. Mary Schooling, Chris Wallace, Qingyuan Zhao, and George Davey Smith. Mendelian randomization. *Nature Reviews Methods Primers*, 2(1):6, 2022.

Enrique F Schisterman, Robert M Silver, Neil J Perkins, Sunni L Mumford, Brian W Whitcomb, Joseph B Stanford, Laurie L Lesher, David Faraggi, Jean Wactawski-Wende, Richard W Browne, Janet M Townsend, Mark White, Anne M Lynch, and Noya Galai. A randomised trial to evaluate the effects of low-dose aspirin in gestation and reproduction: design and baseline characteristics. *Paediatr Perinat Epidemiol*, 27(6):598–609, Nov 2013. DOI: 10.1111/ppe.12088.

Daniel Schwartz and Joseph Lellouch. Explanatory and pragmatic attitudes in therapeutical trials. *Journal of Chronic Diseases*, 20(8):637–648, 1967.

Michaël Schwarzinger, Bruce G Pollock, Omer S M Hasan, Carole Dufouil, and Jürgen Rehm. Contribution of alcohol use disorders to the burden of dementia in france 2008-13: a nationwide retrospective cohort study. *Lancet Public Health*, 3(3):e124–e132, 2018.

Nicholas J. Timpson, Roger Harbord, George Davey Smith, Jeppe Zacho, Anne Tybjærg-Hansen, and Børge G. Nordestgaard. Does Greater Adiposity

Increase Blood Pressure and Hypertension Risk? *Hypertension*, 54(1): 84–90, 2009.

Simon L. Turner, Amalia Karahalios, Andrew B. Forbes, Monica Taljaard, Jeremy M. Grimshaw, Allen C. Cheng, Lisa Bero, and Joanne E. McKenzie. Design characteristics and statistical methods used in interrupted time series studies evaluating public health interventions: A review. *Journal of Clinical Epidemiology*, 122:1–11, 2020.

Kirsten Waller, Shanna H. Swan, Gerald DeLorenze, and Barbara Hopkins. Trihalomethanes in drinking water and spontaneous abortion. *Epidemiology*, 9(2), 1998.

Kazuki Yoshida, Daniel H. Solomon, and Seoyoung C. Kim. Active-comparator design and new-user design in observational studies. *Nature reviews. Rheumatology*, 11(7):437–441, July 2015. ISSN 1759-4790. DOI: 10.1038/nrrheum.2015.30.