Hi, my name's Nick Seewald, I'm a Ph.D. candidate in statistics at the University of Michigan, and I'll be talking about sample size and timepoint tradeoffs for comparing dynamic treatment regimens in a longitudinal SMART. This work is joint with my advisor, Danny Almirall.

I want to start with a motivating example from the field of addiction research. It's well-known that it's hard to engage individuals with alcohol- or cocaine-related substance use disorders in treatments. These interventions are effective, so clinicians really want these patients to re-engage with them after they've disengaged. How should they do that? Should the strategy be to try and get them back into the original treatment, or should they give the patient some autonomy and offer them a choice of treatment modality?

For some people, these re-engagement strategies will work, but others will continue to not engage. Obviously we'd like to get them into treatment, so we want to follow-up with something for those folks who are continued non-engagers.

So this is a question about a *sequence* of treatments that is adapting to the individual over time, and this is often how clinical decision-making works. A *dynamic treatment regimen* operationalizes this clinical decision-making process by recommending specific treatments to certain subsets of individuals at specific times. Here's an example dynamic treatment regimen:

- Initially, this dynamic treatment regimen recommends that patients receive an intervention we're abbreviating as "MI-IOP", which stands for motivational interviews with a focus on an intensive outpatient program. This is basically two phone calls geared toward getting the patient back into their original treatment program.
- Then, if the participant sufficiently engages in the program according to some pre-specified criteria, the dynamic treatment regimen recommends no further contact for that individual: if it ain't broke, don't fix it.
- However, if the patient does not meet the criteria for sufficient engagement, the dynamic treatment regimen recommends a second intervention abbreviated "MI-PC", which stands for motivational interview with patient choice. Here, the participant is given a second set of phone calls, this time focusing on engaging them in a substance use treatment modality of their choosing.
- Notice that the dynamic treatment regimen makes recommendations for *both* engagers and continued non-engagers: it tailors subsequent treatment according to engagement status.

We can address scientific questions about how to construct a high-quality dynamic treatment regimen using a sequential multiple-assignment randomized trial, or SMART. A SMART is just one type of randomized trial that can answer these questions. The key feature of a SMART is that some or all participants are randomized more than once.

Let's take a look at an example. This is a study called ENGAGE. The PI is Jim McKay, who's at UPenn. In the trial, 500 individuals with an alcohol- and or cocaine-related substance use disorder who did not sufficiently engage in the first 8 weeks of an intensive outpatient program. These folks were randomized between MI-IOP and MI-PC, then followed for another 8 weeks, at which point they were determined to have either sufficiently engaged, or continued to not engage. This was determined using prespecified attendance criteria. Everyone who engaged was given no further contact (but still followed for research outcomes). All continued non-engagers were randomized a second time, between MI-PC and no further contact. Everyone was followed for a total of 24 weeks. Over the course of the study, a (for our purposes) continuous outcome called "treatment readiness", a measure of a participant's willingness and ability to engage in treatment, was measured at baseline and weeks 4, 8, 12, and 24.

You may have noticed that there are dynamic treatment regimens *embedded* in the SMART. There are four such embedded regimens. Here's the one we saw earlier, which recommends MI-IOP initially, then no further contact for engagers and MI-PC for continued non-engagers. Individuals who end up in subgroups A and B at the end of the trial are "consistent" with this regimen. Similarly, here's another embedded regimen, and another, and another.

There are two things I want to point out here: The first is that engagers are consistent with two embedded DTRs, whereas non-engagers are only consistent with one. Notice that subgroup D lights up in both blue *and* yellow; subgroups E and F light up only once. The second point is that engagers are only randomized once, whereas non-engagers are randomized twice. Both things need to be accounted for in our analysis of the data which arise from this trial.

Our goal is to develop a sample size formula for the comparison of two embedded DTRs at the end of the study using a longitudinal outcome collected at an arbitrary number of timepoints.

Because sample size is chosen for a specific analysis, we need to talk about modeling. Here's one way to model data in a longitudinal SMART. This is a piecewise-linear model which reflects the sequential nature of treatment delivery in a SMART. At baseline, we don't expect individuals to differ by assigned treatment, so there's one mean. Between times 0 and 8, there are two groups: one for each first-stage treatment. After time 8, there are now four groups: one for each embedded dynamic treatment regimen. Notice that this model is marginal over engagement status, since the regimens recommend treatment for both engagers *and* continued non-engagers.

Remember that our primary aim is a comparison of the purple line to the blue line at week 24.

So how do we fit the model? We use GEE-*type* estimating equations. You'll see why I say GEE *type* in just a second. If we start down at the end of the second line, we've got our usual vector of residuals, a working covariance matrix, and the Jacobian of our model, just like usual GEE. The top line is where things get slightly different. We're now adding an inverse-probability weight to account for the fact that engagers are randomized once, whereas non-engagers are randomized twice. The weight also includes an indicator that specifies whether a particular individual is consistent with a given embedded dynamic treatment regimen. We're now also summing over those embedded regimens, to capture the fact that engagers are consistent with two, and non-engagers are only consistent with one DTR each.

So. Let's more formally state our goal. We want to develop a tractable sample size formula for the test of whether the expected difference in potential outcomes between two embedded DTRs at the end of the study is zero, against some fixed alternative, Δ. The superscript here denotes the potential outcome under a particular DTR. Using the model we saw earlier, we can write this estimand as a linear combination of our model parameters, which are asymptotically normal. So, we can just use an asymptotic Z test for this comparison, using the sandwich variance of the regression parameters beta-hat. The challenge here is to get a tractable upper bound on that variance.

Under some mild working assumptions and an exchangeable within-person covariance structure with constant variance across time and regimen, this is our sample size formula. It can be decomposed into three parts. The first is the standard sample size formula for a two-arm RCT. δ is the target standardized effect size, determined in part by the fixed alternative hypothesis from earlier. $1-\gamma$ is the target power. The second term is an inflation factor associated with our SMART design. R, here, is an indicator of whether an individual engaged with first-stage treatment. This term ranges from 1 – 2, and can be thought of as trading off between a 2-arm trial (if everyone's an engager, nobody's re-randomized), and a 4-arm trial (if everyone's a non-engager, everyone's re-randomized). Finally, we have a deflation factor associated with the longitudinal outcome. Note that this is a *de*flation factor, since we have *within*-person correlation and we're making a *between*-groups comparison. The deflation factor depends on the strength of the within-person correlation, ρ, the number of measurement occasions in stage 2, T_2, and the total number of measurement occasions in the trial, T. Our long-term goal for this project is to understand how to trade off between the sample size, T_2, and T in order to maximize power subject to a budget constraint.

In the special case of 3 timepoints: one at baseline, one at the end of the first stage, and one at the end of the second stage, our deflation factor reduces to $1-\rho^2$. This is exactly the deflation factor you'd see in a two-arm pre-post RCT when you don't model a difference in means at baseline. This work is published in *Statistical Methods in Medical Research*.

One strategy for generalizing the number of timepoints is to add measurements equally in both stages of the SMART. Let's see what happens to our sample size requirement using this strategy.

Here's a plot of the deflation factor on the y axis versus the within-person correlation ρ on the x axis. A value of 1 on the y axis corresponds to no deflation in the sample size: this is a multiplicative "effect" here. As we add timepoints in both stages so that the total number of measurement increases (higher values of T correspond to lighter lines), we can see meaningful decreases in the deflation factor: more measurement occasions gets us more power, but the impact lessens for large values of ρ.

Now we might ask what happens when we distribute timepoints unequally across the stages of the trial. So we could put more in stage 1 than in stage 2, or more in stage 2 than stage 1. Let's see what happens when, with a fixed number of timepoints in the whole study, we move measurements between stage 1 and stage 2.

So here we're looking at a total 7 timepoints in the whole study, and increasing T_2, the number of timepoints in stage 2 of the SMART. Again, the y-axis on these plots is the deflation factor, so 1 corresponds with no benefit in terms of sample size. On the left, the x axis is again ρ, and this dashed line is $1-\rho^2$, the deflation factor for a 3-timepoint SMART. As we shift more measurements into the second stage, we see noticeable drops in the deflation factor (so, more power), for small ρ. For larger values of ρ, it doesn't matter much. We can see that more clearly in the plot on the right. Here, the x-axis is now T_2 and the lines represent different values of ρ. As ρ increases, the lines get flatter, indicating that shifting timepoints into stage 2 has less impact on power. Curiously, there's some non-monotonicity happening here. When T_2 is large, small values of ρ slightly hurt you. We don't quite have intuition for that yet.

Wrapping up, this is a work in progress. We're in the process of building a user-friendly sample size tool that can abstract away some complexity in the functional form of the deflation factor for non-statistical folks, and we're still working on developing guidance for how to balance sample size and timepoints subject to a cost-constraint. We’re also still thinking about that non-monotonicity in the relationship between the deflation factor and ρ for large values of T_2.

As I mentioned earlier, the special case of 3 timepoints has appeared in Statistical Methods in Medical Research; the author accepted manuscript is also on the ArXiv. Finally, I’d like to acknowledge funding from the NIH and the Institute of Education Sciences, and you can find slides and other resources on my website, nickseewald.com. Thanks so much for stopping by, and enjoy the rest of JSM.