We introduce Dr. Alex Kaizer. Dr. Kaizer is an assistant professor in the Department of Biostatistics and Informatics, and he’s a faculty member in the Center for Innovative Design and Analysis at the University of Colorado Medical Campus. He’s passionate about translational research and the development of normal and clinical trial designs that are more efficient and factually utilize available resources including past trials and past studies. Dr. Kaizer strives to translate (indistinct) topics into understandable material that is more than just a mask and something we can appropriate and utilize in our daily lives and research. Now let’s welcome Dr. Kaizer.

Thank you Wayne. So apologies for my own technical difficulties today, but I’m going to be presenting on this idea of a sequential basket trial design based on multi-source exchangeability with predictive probability monitoring. And that is admittedly quite the mouthful and I’m hoping throughout this presentation to break down each of these concepts and ideas building upon them sort of until we have this
cumulative effect that represents this title today.

Before jumping into everything though, I do wanna make a few acknowledgements. This paper was actually published just at the end of this past summer in PLOS ONE, and so if you’re interested in more of the technical details or additional simulation examples and things beyond what I present today, I include this paper here and we’ll also have it up again at the very end of my talk.

Also acknowledgement to Dr. Nan Chen who helped with some of the initial coding of some of these methods and approaches.

So to set the context for my seminar today, I want to think here about this move towards precision medicine generally, but especially in the context of oncology. And so within oncology, like many other disciplines, when we design research studies, we often design these for a particular histology or indication or a disease. So for example, we might say, "Well, I have a treatment or intervention which I hope or think will work in lung cancer, therefore I’m going to design
and enroll in the study for lung cancer.”

Now this represents a very standard way that we do clinical trial design where we try to really rigorously define and limitedly define what our scope is.

Now within oncology, we’ve had some exciting scientific developments over the past few decades. So now instead of seeing cancer as just based on the site like you have a lung cancer or a prostate cancer, we actually have identified that we can partition cancers into many small molecular subtypes.

And further, we’ve actually been able to leverage this information by being able to say that what we thought of as a holistic lung cancer isn’t just one type of disease, we can actually develop therapies that we hope to target some of these differences in genetic alterations.

And this really gets to that idea of precision medicine that instead of throwing a treatment at someone where we think it should work or it has worked in some people on average, hopefully we can really target the intervention based off of some signal or some indication like a biomarker or a genotype.
that we actually hope could respond more ideally
to that intervention.
Now what’s really interesting about this as well is that there could be a potential for heterogeneity in this treatment benefit by indication.
And what I mean by that is once we’ve identified that there’s these different genetic alterations, we’ve actually discovered that these alterations aren’t necessarily unique to one site of cancer. For example, we may identify a genetic alteration in the lung that also is present in the prostate, liver, and kidney in some of those types of cancer. Now the challenge here though is that even though we have the same driver hypothetically based on our clinical or scientific hypothesis of that potential benefit for a treatment we’ve designed to address it, there’s still may be important differences that we don’t know about or have yet to account for based off of each site.
So what may have worked actually really well in the lung for one given mutation, even for that same mutation, let’s say present in the liver, may not work as well.
And that’s that idea of heterogeneity and treatment benefit. That we can have different levels of response across different sites or groups of individuals. Now the cool thing I think here from the statistical perspective is that the scientific and clinical advancements have also led to the revolution and statistical design challenges and approaches. And of course that’s the sweet spot that I work at. I know many of you and especially students are training and studying to work in this area to collaborate with scientific and clinical researchers and leaders to translate those results in statistically meaningful ways and to potentially design trials or studies that really target these questions and hypotheses.

Now specifically in this talk today, I’m going to focus on this idea of a master protocol design or evolution. And these provide a flexible approach to the design of trials with multiple indications, but they do have their own unique challenges that I’m gonna highlight a few of here in a second. But there are a variety of master protocols out there.
in case you’ve heard some of these buzzwords. I’ll be focusing on basket trials today, but you may have also heard of things like umbrella trials or even more generally platform trial designs. And so one example of what this looks like here is this is a graphic from a paper in the New England Journal by Dr. Woodcock and LaVange, Dr. Woodcock being a clinician, and Dr. Lisa LaVange being a past president of The American Statistical Association, where they actually tried to put to rest some of the confusion surrounding some of these design types because it turns out, up until 2017 when we discussed these designs across even statistical communities and with clinical researchers, we tend to use these terms fairly interchangeably even though we are really getting at very different concepts. So for example, in the top here we have this idea of an umbrella trial and this is really the context of a single disease like lung cancer, but we actually then will screen for those genetic alterations
and have different therapies that we’re trying to target a different biomarker or genetic alteration for. This contrasts to what we’re focusing on today below of a basket trial, we actually have different diseases or indications, but they share a common target or genetic alteration which we wish to target. And in this sense we can think of it potentially as them sharing a basket or sharing a sort of that commonality there. Now, this is a fairly broad general idea of these designs. And so I think for the sake of what we’re gonna talk about today and some of the statistical considerations that can be helpful to do a bit of a oversimplification of what a design might look like here. And so on the slide that I’ve presented, I have this kind of naive graphic of actual baskets and we’re going to assume that in each column we have a different indication or site of cancer that has that common genetic alteration. So for example, basket one may represent the lung, basket two may represent the liver and so on. Now when we’re in the case of designing or the design stage of a study,
we tend to make oversimplifying assumptions to address these potential calculations for power, sample size, and quantities that we’re usually interested in for study design. So here on this graph, we are gonna make an assumption that there’s only two possible responses in this planning stage. One is that the baskets have no response or a null basket, and the other case would be an alternative response where there is some hopeful benefit to the treatment. Now, one of the challenges I think with basket trial design is that we do have to make this assumption of what is our null hypothesis or response? What’s our alternative hypothesis or response? We really only have to do that for one configuration because we have two arms.
just by having five baskets in a study and many actual trials that are implemented at far more baskets, we actually see a range of just six possible binary combinations of the basket works or it doesn’t work, ranging from at the extremes a global null where unfortunately the treatment does not work varying from at the extremes a global null where the treatment does not work, or the drug actually works across all baskets. There is this homogenous actually response in a positive direction for the sort of clinical outcome. More realistically, we actually will probably encounter something that we see falls in the middle here, scenarios two through five, where there’s some mixture of baskets that actually do show a response and some that for whatever reason we might not know yet, it just doesn’t appear to have any effect and is a null response. So this can make it challenging for some of the considerations of what analysis strategy you plan to use in practice. And so to just, at a high level, highlight some of these challenges
before we jump into the methods for today’s talk.

In practice, each of these baskets within trial often have what we call a small and or small sample size for each of those indications.

It turns out once we actually have this idea of precision medicine and we can be fairly precise for who counts for a trial, we actually have a much smaller potential sample or population to enroll.

This means that even though we might have a treatment that works really well, it can be challenging to find individuals who qualify or are eligible to enroll or they may have competing trials or demands for other studies or care to consider.

As I’ve also alluded to earlier the challenge, we also have this potential for indication or subgroup heterogeneity and that may be likely.

In other words, we might not expect the same response across all those baskets. And that gets back to the previous graphic on that last slide where we might have something like two null baskets.
and three alternative baskets. And that can make it really challenging in the presence of a small n to determine how do we appropriately analyze that data so we capture the potentially applications baskets and can move those forward so patients benefit while not carrying forward null baskets where there is no response for those patients. Statistically speaking, we also have these ideas of operating characteristics and in the context of a trial, what we mean by that is things like power and type one error and I just have additional considerations with respect to how do we summarize these? Do we summarize them within each basket or each column on that graphic on the previous slide, essentially treating it as a bunch of standalone independent one arm trials just under one overall study design or idea? Or do we try to account for the fact that we have five baskets enrolling like on the graphic before and we might wanna consider something like a family wise type one error rate where any false positive would be a negative outcome.
if we're trying to correctly predict or identify associations?

Now the focus of today's talk, and I could talk about these other points till the cows come home, but I'm gonna focus today on depending on that research stage we're at, if it's a phase one, two or three trial, we may wish to terminate early for some reason like efficacy or futility. And specifically for time today, I'm gonna focus on the idea of stopping for futility where we don't wanna keep enrolling baskets that are poorly performing both for ethical reasons. In other words, patients may benefit from other trials or treatments that are out there and we don't wanna subject them to treatments that have no benefit. But also from a resource consideration perspective. You can imagine that running a study or trial is expensive and can be complicated. And especially if we're doing something like a basket trial where we're having to enroll across multiple baskets, it may be ideal to be able to drop baskets early on
that don’t show promise so we can reallocate those resources to either different studies, research projects, or trials that we’re trying to implement or run.

So the motivation for today’s talk building off of these ideas is that I want to demonstrate that a design that’s very popular called Simon’s two-stage design is generally speaking suboptimal compared to the multitude of alternative methods and designs that are out there.

And then this is especially true in our context of a basket trial where within the single study we actually are simultaneously enrolling multiple one arm trials in our case today.

Then the second point I’d like to highlight is we can identify when methods for sharing information across baskets could be beneficial to further improve the efficiency of our clinical trials.

And so to highlight this, I wanna first just build us through and sort of illustrate or introduce these designs and the general concepts behind them because I know if you don’t work in this space it may be sort of just ideas vaguely.

So I wanna start with the Simon two-stage design,
that comparator that people are commonly using.

So Richard Simon, and this is back in 1989,

introduced what he called optimal two-stage designs

for phase two clinical trials.

And this was specifically in the context

that we're focusing on today for a one sample trial

to evaluate the success of a binary outcome.

So for oncology we might think of this as a yes no outcome

for is there a reduction in tumor size

or a survival to some predefined time point.

Now specifically what Dr. Simon was motivated by

was the stage-two trials

as it says in the title of his paper,

and just to kind of give a common lay of the land for everyone,

the purpose generally speaking of a phase two trial

is to identify if the intervention warrants further development

while collecting additional safety data.

Generally speaking,

we will have already completed what we call a phase one trial where we collect preliminary safety data

to make sure that the drug is not toxic

or at least has expected side effects

that we are willing to tolerate for that

potential gain in efficacy.
And then in phase two here we’re actually trying to say, “You know, is there some benefit? Is it worth potentially moving this drug on either for approval or some larger confirmatory study to identify if it truly works or doesn’t?”

Now the motivation for Dr. Simon is that we would like to terminate studies earlier, for both ethical and resource considerations that they appear futile. In other words, it’s not a great use of our resources and we should try in some rigorous statistical way to address this. If you do go back and look at Simon’s 1989 paper or you just Google this and there’s various calculators that people have put out there, there are two flavors of this design that exist from this original paper. One is an optimal and one is called a minimax design. Within clinical trials, once we introduce this idea of stopping early potentially or have the chance to stop early based on our data, we now have this idea that there’s this expected sample size.
because we could enroll the entire sample size that we planned for or we could potentially stop early.

And since we could stop early or go the whole way and we don’t know what our choice will be until we actually collect the data and do the study,

we now have sample size of the random variable, something that we can calculate an expectation or an average for.

And so Simon’s optimal design tries to minimize what that average sample size might be in theory.

In contrast, the minimax design tries to minimize whatever that largest sample size would be if we didn’t stop early.

So if we kept enrolling and we never stopped at any of our interim looks,

how much data would we need to collect until we choose a design that minimizes that at the expense of potentially stopping early?

I think this is most helpful to see the sort of elegance of this design and why it’s I think so popular by just introducing example that will also motivate our simulations here that we’re gonna talk about in a minute.

We’re gonna consider a study where
The null response rate is 10%. And we're going to consider a target alternative response rate of 30%. So this isn't a situation where we're looking for necessarily a curative drug, but something that does show what we think of as a clinically meaningful benefit from 10 to 30%.

Let's say survival or tumor response. Now if we have these two parameters we wanna do a Simon two-stage minimax design to minimize that maximum possible sample size. We would enroll 16 participants and would terminate the trial or this study arm for futility if one or fewer responses are observed.

We also wish to target a 90% power to detect that treatment of 30% if it truly exists. So we put all of this into our calculator and we turn that statistical crank. What we see is that it gives us this approach where in stage one we would enroll 16 participants and we would terminate the trial or this study arm for futility if one or fewer responses are observed.
Now if we observe two or more responses, we would continue enrollment to the overall maximum sample size that we plan for 25 in the second stage. And at this point if four or fewer responses are observed, no further investigation is warranted or we can think of this as a situation where our P value would be larger than our defined alpha 0.1. Now, the nice thing here is that it is quite simple. In fact, after we trim that statistical crank and we have this decision rule, you in theory don’t even need a statistician because you can count the number of responses for your binary outcome on your hand and determine should I stop early, should I continue? And if I continue, do I have some benefit potentially that says it’s worth either doing a future study or I did a statistical test, would find that the P value meets my threshold. I set for significance. Now, of course, it wouldn’t be a great talk if I stopped there and said, "You know, this is everything. It’s perfect. There’s nothing to change."

There are some potential limitations...
and of course some solutions I think that we could address in this talk. The first thing to note is that this is extremely restrictive in when it could terminate. It may continue to the maximum sample size even if a null effect is present. And we’re gonna see this come to fruition in the simulation studies, but it’s worth noting here it only looks once. It’s a two stage design. And depending on the criteria you plug in, it might not look for quite some time. 16 out of 25 total participants enrolled is still a pretty large sample size relative to where we expect to be. One solution that we could look at and that I’m going to propose today is that we could use Bayesian methods instead for more frequent interim monitoring. And this could use quantities that we think of as the posterior or the predictive probabilities of our data. Another limitation that we wish to address as well is that in designs like a basket trial that have multiple indications or multiple arms that have the same entry criteria, Simon’s two-stage design is going to
fail to take advantage of the potential what we call exchange ability across baskets. In other words, if baskets appear to have the same response, whether it’s let’s say that null or that alternative response, it would be great if we could informatively pull them together into meta subgroups so we can increase the sample size and start to address that challenge of the small n that I mentioned earlier for these basket trial designs. And specifically today we’re going to examine the use of what we call multi-source exchangeability models to share information across baskets when appropriate. And I’ll walk through a very high level sort of conceptual idea of what these models and how they work and what they look like. Before we get into that though, I wanna just briefly mention the idea of posterior and predictive probabilities and give some definitions here so we can conceptually envision what we mean especially if you haven’t had the chance to work with a lot of patient methods, this can help give us an idea
of some of the analogs to maybe a frequentist approach or what we’re trying to do here that you may be familiar with. Now I will mention, I’m not the first person to propose looking at Bayesian interim stopping rules. I have a couple citations here by Dmitrienko and Wang and Saville et al and they do a lot of extensive work in addition to hundreds of other papers considering Bayesian interim monitoring. But specifically to motivate this we have these two concepts that commonly come up in Bayesian analysis, a posterior probability or a predictive probability. The posterior probability is very much analogous to kinda like a P value in a frequent significance. It says, "Based on the posterior distribution we arrive at through a Bayesian analysis, we’re gonna calculate the probability that our proportion exceeds the null response rate we wish to beat.” So in our case, we’re basically saying, "What’s the probability based on our data and a prior we’ve given that the response is 10% or higher.” So this covers a lot of ground
'cause anything you know from 10.1 up to 100% would meet this criteria being better than 10%.

But it does quantify, based on the evidence we’ve observed so far, how the data suggests the benefit may be with respect to that null. So in the case of let’s say an interim look for futility at the data, we could say, if we just use Simon’s two-stage design as our motivating ground to consider, we might say, "Okay, we have 16 people so far, what’s the probability based on these 16 people that I could actually say there’s no chance or limited chance I’m going to detect something in the trial here based on the data I’ve seen so far?"

Now the challenge here is that it is based on off the data we’ve seen so far and it doesn’t take into account the fact that we still have another nine potential participants to enroll to get to that maximum sample size of 25. That’s where this idea of what we call a predictive probability comes in. We’re considering our accumulated data and the priors we’ve specified in our Bayesian context, it’s the probability that we will have observed
a significant result if we’ve met and enrolled up to our maximum sample size. In other words, I think it’s a very natural place to be for interim monitoring because it says based on the data I’ve seen so far, i.e the posterior probability, if I use that to help identify what are likely futures to observe or likely sample sizes I will continue enrolling to get to that maximum of 25, what’s the probability at the end of the day when I do hit that sample size of 25, I will have a significant conclusion? And if it’s a really low predictive probability, if I say there’s only a 5% chance of you actually declaring significance if you keep enrolling participants, that can be really informative both statistically and for clinical partners to say it doesn’t seem very likely that we’re gonna hit our target. That being said, a lot of people are very happy to continue trials going with low chances or low probability because you’re saying there’s still a chance I may detect something that could be significant enough worth.
So we’ll see that across a range of these thresholds, the performance of these models may change.

Now this brings us to a brief recap of sort of our motivation. I just spent a few minutes introducing that popular Simon two-stage design, the idea behind it, what it might look like in practice, as well as some alternatives with the Bayesian flare.

The next part I wanna briefly address is that we can also now look at this idea of sharing information across baskets to further improve that trial efficiency ’cause so far both Simon’s design and the just using a posterior predictive probability for an interim monitoring will still treat each basket as its own little one arm trial.

Now specifically today I’m gonna focus on this idea we call multi-source exchangeability models or MEMs. This is a general Bayesian framework to enable the incorporation of independent sources of supplemental information and its original work that I developed during my dissertation at the University of Minnesota.

In this case,
the amount of borrowing is determined by the exchange ability of our data, which in our context is really how equivalent are the response rates? If two baskets have the exact same response rate, we may think that there's a higher probability that the true underlying population we are trying to estimate are truly exchangeable. We wish to combine that data as much as we possibly can.

First is if again we see something that is like a 10% response rate for one basket and a 30% response rate for another basket, we likely don't want to combine that data because those are not very equivalent response rates. In fact, we seem to have identified two different subgroups and performances in those two baskets. One of the advantages of MEMs relative to a host of other statistical methods that are out there that include things like power priors, commensurate priors, meta analytic priors, and so forth, is that we've been able to demonstrate that in their most basic iteration without any extra bells or whistles, MEMs are able to actually account for this heterogeneity.
across different potential response rates and appropriately down weight non-changeable sources. Whereas we show through simulation and earlier work some of these other methods without newer advancements to them actually either naively pull everything together even if there’s non-changeable groups or they’re afraid of the sort of presence of non-change ability and if anything seems amiss, they quickly go to an independence analysis that doesn’t leverage this potential sharing of information across meta subgroups that are exchangeable. Now again, I don’t wanna get too much into the weeds of the math behind the MEMs, but I will have a few formulas in a couple slides but I do think it’s helpful to conceptualize it with graphics. And so here I just want to illustrate a very simplified case where we’re gonna assume that we have a three basket trial and for the sake of doing an analysis with MEMs, I think it’s helpful to also think of it as we’re looking at the perspective of the analysis from one particular basket.
So here on this slide here we see that we have this theta P circle in the middle and that’s the parameter or parameters of interest we wish to estimate. In our case, that would be that binary outcome in each basket. Now, for this graphic we’re using each of these circles here to represent a different data source. We’re gonna say Y sub P is that primary basket that we’re interested in or the perspective we’re looking at for this example and Y sub one and Y sub two are two of the other baskets enrolled within the trial. Now a standard analysis without any information sharing across baskets would only have a data pooled from the observed data. I mean this is sort of the unexciting or unsurprising analysis where we basically are analyzing the data we have for the one basket that actually represents that group. However, we could imagine if we wish to pool together data from these other sources, we have different ways we could add arrows to this figure
to represent different combinations of these groups.

And this brings us to the multi-source exchangeability framework. So we see here on this slide, I now of a graphic showing four different combinations of exchangeability when we have these two other baskets.

And from top left to the bottom left in sort of a clockwise fashion, we see that making different assumptions from that standard analysis with no borrowing.

So it is possible that none of our data sources are exchangeable and we should be doing an analysis that doesn’t share information.

On the right hand side that we might envision well maybe the first basket or Y1 is exchangeable.

So we wanna pull that with Y2 or excuse me with Yp,

but Y2 is not.

In the bottom right, this capital omega two, we actually assume that Y2 is exchangeable but Y1 is not.

And in the bottom left we assume in this case that all the data is exchangeable.
and we should just pool it all together.

So at this stage we’ve actually proposed all the configurations we can pairwise of combining these different data sources with $Y_{sub \;P}$.

And we know that these are fitting four now different models based off of the data because for example in the top left, that standard analysis,

there is no extra information from those other baskets versus like in the bottom left,

we basically have combined everything and we think there’s some common effect.

Now this leads to two challenges on its own if we just stopped here with the framework.

One would be that we’d have this idea of maybe cherry picking or trying to pick whichever combination best suits your prior hypotheses clinically.

And so that would be a big no-go.

We don’t like cherry picking or fishing for things like $P$ values or significance in our statistical analyses.

The other challenge also is that all of these configurations are just assumptions of how we could combine data but we know underlying everything in the population is that true assumption of exchange ability of
Are these baskets or groups truly combinable or not?

And we’re just approximating that with our sample.

And so right now if we have four separate models

and potentially four separate conclusions,

we need some way of combining these models to make inference.

And in this case we propose leveraging a Bayesian model averaging framework

where we calculate in this case and in our formulas here,

the queues represent a posterior distribution where I’ve drawn this little arrow

and I’m underlining right now,

that reflects each square’s configuration of exchange ability for our estimates.

And through this process we estimate these lower case omega model weights

that tries to estimate the appropriateness of exchangeability with the ultimate goal of having a average posterior that we can use for statistical inference to draw a conclusion about the potential efficacy or lack thereof of a treatment.

Now very briefly, because this is a Bayesian model averaging framework,
just one of the few formulas I have in the presentation, we just see over here that we have the way we calculate these posterior model weights as the prior on each model multiplied by an integrated marginal likelihood. Essentially, we can think of that as saying based off of that square we saw on the previous slide and combining those different data sources, what is that estimate of the effect with those different combinations? One unique thing about the MEM framework that differs from Bayesian model averaging though is that we actually specify priors with respect to these sources. And in the case of this example with only two supplemental like sources for our graphic, it’s not a great cost savings, but we can imagine that if we have more and more sources, there’s actually two to the \( P \) if \( P \)’s the number of sources, combinations of exchange ability that we have to consider and model. And that quickly can become overwhelming if we have multiple sources that we have to define for each one of those squares,
what’s my prior that each combination of exchangeability is potentially true.

Versus if we define it with respect to the source,

we now go from two to the P priors to just P priors we have to specify for exchangeability.

A few more notes about this idea here and just really zooming in on what we’re gonna focus on for today’s presentation.

We have developed both fully and empirically Bayesian prior approaches here,

fully Bayesian meaning that it is defined a priori and is agnostic to the data you’ve collected,

empirically Bayesian meaning we leverage the data we’ve collected to help inform that prior for what we’ve observed.

Specifically there is a what we call a non constrained, or naive, empirically based prior where we would look through all of those growths we had and we would say, "Whichever one of these maximizes the integrated marginal likelihood that’s the correct configuration and we’re gonna put all of our eggs into that basket."

Or 100% of the probability there and that’s the only model we use for analysis.
We know, generally speaking, since we went to all the work to defining all of these different combinations of exchangeability and that it’s based off of samples, potentially small samples, that this can be a very strong assumption. And so we can also modify this prior to what we call a constrained EB prior, where instead of just giving everyone of those model sources in that MEM that maximizes the likelihood 100% weight, we instead give it a weight of what we’re calling just B. This is our hyper prior value where if it’s a value of zero or up to one, it’ll control the amount of borrowing and allow other nested models of exchangeability to also be potentially considered for analysis. So for example, if we do set a value of one, that actually replicates the non constrained EB prior and really aggressively borrows from one specific model. At the other extreme here, if we set a value of zero, we essentially recreate an independent analysis like assign a two stage design or just using those Bayesian methods for futility monitoring.
that doesn’t share information. And then any value in between gives a little more granularity or control over the amount of borrowing.

So with that background behind us, I’m gonna introduce the simulation stuff and then present results for a couple key operating characteristics for our trial. In this case, we’re going to assume for our simulations that we have a basket trial with 10 different baskets or indications. So again, that’s 10 different types of cancer that have enrolled that all have the same genetic mutation that we think is targeted by the therapy of interest.

Like we had before, we’re going to assume a null response $P_{knot}$ of 0.1 or 10%. And an alternative response rate of 30% or $P_1$ here.

We are gonna compare then three different designs that we just spent some time introducing and outlining. The first is a Simon minimax two-stage design using that exact set up that we had before where we will enroll 16 people, determine if we have one or fewer observations of success.

If so, stop the trial. If not, continue on.
In the second case, we’re going to implement a Bayesian design that uses predictive probability monitoring but we don’t use any information sharing just to illustrate that we can at least potentially improve upon the frequency in use of a interim monitoring above a single look from the Simon minimax design.

And then the third design will add another layer of complexity where we will try to share information across baskets that have what we estimate to be exchangeable subgroups.

One thing to note here is that we are setting this hyper parameter value $B$ at 0.1. This is a fairly conservative value and admittedly for this design we actually did not calibrate specifically for the amount of borrowing to be 0.1.

This is actually based off of some other prior work we’ve done and published on basket trials that just showed that in the case of an empirically Bayesian prior for MEMs, this actually allows information sharing in cases where there’s a high degree of exchangeability and low heterogeneity and down leap it in cases where we might be
a little more uncertain, so it’s a little more conservative but we’ll see in the simulation results there are some potential benefits.

For each of the scenarios we’re gonna look at today, we will generate a thousand trials with a maximum sample size of 25 per basket. We’re gonna look at two cases, there’s a few other in the paper but we’re gonna focus on first the global scenario where all the baskets are either null or all 10 baskets have some meaningful effect. And this is the setting where information sharing methods like meds really should outperform anything else because everything is truly exchangeable and everything could naively be pooled together because we’re simulating them to have the same response. We’ll then look at what happens if we actually have a mixed scenario, which I think is actually more indicative of what’s happened in practice with some of the published basket trials and clinically what we’ve seen from applications of these types of designs. Specifically here, we’re gonna look at the case where
there are eight null baskets and two alternative baskets.

A few other points just to highlight here.

We're going to assume a beta 0.5 0.5 prior for our Bayesian models.

This essentially for a binary outcome can be thought of as adding half of a response and half of a lack of a response to our observed data.

We're going to look at the most extreme dream Bayesian case of doing utility monitoring or any type of interim monitoring continually. So after every single participant's enrolled we will do a calculation and determine if we should stop the trial.

We will then look at the effect of this choice across a range of predictive probability thresholds ranging from 0%, meaning we wouldn't stop early at all, up to 50% saying if there's anything less than a 50% chance I'll find success, I wanna stop that trial.

And then finally it's worth noting we're actually also completely disregarding calibration for this interim monitoring.

And so what we're gonna do is we're gonna calibrate our decision rules for the posterior probability at the end of the trial.
based off of a global scenario where we think it’s ideal to share information and we’re all not gonna account for the fact that we’re doing interim looks at the data. Part of the question here was if we truly do all these assumptions and we do sort of the most naive thing, how badly do we actually do? Like is there enough reason to fear the results if we don’t correctly calibrate for everything here?

I’m gonna paint some pictures here building from the simpler Simon design to our more complex Bayesian designs and then with information sharing just to illustrate three different properties. I’m gonna go fairly quickly ’cause I know that you all have to vacate the classroom in about 10 minutes.

For the global scenario that we’re looking at here, the like rate lines are going to represent the alternative basket scenario. So all, in this case, all 10 null baskets. Here we see we plan for 90% power and Simon’s design appropriately achieved that rejection rate of 90%. Likewise, the lines at the bottom here, these black lines,
Here are the global null scenario and we see that it achieves a 10% rejection rate.

Now, this is a flat line here because again Simon’s design is agnostic to things like the predictive probability.

Now if we do frequent Bayesian monitoring, we see two interesting things here with these new lines.

We see that at the top and the bottom, here I add these circles where the predictive probability threshold is 0%.

This does represent the actual design that would correspond to the actual calibration we did without interim monitoring.

And we see that it is possible with Bayesian approaches to achieve the same frequent operating characteristics.

that we would achieve with something like the Simon design.

We can see though that if we want to do interim monitoring but we didn’t calibrate or think of that in our calculations,

we do see this trade off where we have our alternative baskets having a decreasing power or rejection rate as the aggressiveness of the predictive probability threshold increases.
And likewise the type one error rate or the rejection rate of the marginal baskets also decreases.

Now if we add information sharing to this design, we actually see some encouraging results in this global scenario.

First, it's worth noting that in the case where we actually calibrated for, we actually see an increase in power from 90% to about 97%.

And even when we actually have a 10% predictive probability threshold for interim monitoring, we see that we actually still achieve 90% power with a corresponding reduction in that type one error rate.

Of course, this is with the caveat that this is the ideal setting for sharing information because all of the baskets are truly exchangeable.

Now the rejection rate correlates to something we call that expected sample size.

What is the average sample size we might enroll for each basket of our 10 baskets in the trial?

We see here that in the case of a null basket the Simon design is about 20.

If we do interim monitoring with Bayesian approaches and no information sharing,
obviously if we don’t do any interim looks at
the data,
we have a 0% threshold,
we’re gonna have a sample size of 25 every
single time.
I think what’s encouraging though is that
by looking fairly aggressively we see that our
sample size,
even with a very marginal
or low 5% threshold for futility monitoring,
drops from 20 in the assignment design to
about 15
in the Bayesian design,
the trade-off of course being because we
didn’t calibrate.
We also see a reduction in the sample size
for the alternative baskets.
And if we add that layer of information
sharing,
we actually see that we do slightly better
than the design without information sharing
while attenuating at the top here the effect
our solid gray line has for the alternative
baskets.
Now, briefly tying this together then to the
stopping rate,
which we can kind of infer from those past
results,
we do see that on average the Simon two-
stage design
for the null baskets stopping for futility
is only taking place a little over 50% of the
time
in this simulation.
The advantage here though is that it is very rarely stopping for the alternative baskets.

In our Bayesian approaches, we see that there is an over 80% probability of stopping if it’s a null effect. And this is ideal because we have 10 baskets.

And so these potential savings or effects can compound themselves across these multiple baskets.

We then see that the design adding these solid lines for information sharing do very similarly where again the consequence of not calibrating are attenuated in this circumstance.

Now the thing to note here that everything I presented on these few graphics were with respect to the global scenario, that ideal scenario that I actually don’t think is super realistic in practice.

So we see here, if we do a mixed scenario where we now have calibrated for the global scenarios, we’ve miscalibrated with respect to that.

We’ve also not calibrated for interim looks at the data.

We can actually see that the results for the Simon two-stage in the Bayesian design without information sharing are very similar
to what we saw before.
That’s because they don’t share information.
And so in this case with eight null baskets
into alternative baskets,
they have very similar responses.
This contrasts of course with the MEM approach
or the information sharing approach
where we actually see now
many of these results are actually overlapping
for information sharing and no information sharing.
What this tells us is that even though we miscalibrated
up and down the design,
we are actually able with this more conservative prior
to down weight borrowing
and effectuate similar results
that at lower thresholds for utility monitoring
for example
at 5% can still show potential gains in efficiency relative
to the Simon design that could likely further be improved
with actual calibration.
So just as a reminder,
we demonstrated today
and introduced the idea of Simon’s two-stage design
and some alternative methods to compete with them.
And some just brief discussion and concluding points.
There is no free lunch and this is true regardless of where we are in statistics.
That for example in our designs, besides the fact that we miscalibrated and made it a bit harder of a comparison for our methods, we did try to replicate what people might be doing in practice or the challenge of calibrating these designs into actuality.
Simon’s two-stage design does have a lot of benefits from it’s ideal characteristics that are easy to implement, but it is limited in how often it may stop. Our Bayesian designs, with or without information sharing, can lead to reductions in the expected sample size in the null basket.
Our Bayesian designs, with or without information sharing, can lead to reductions in the expected sample size in the null basket. and further could be improved if we actually incorporate calibration, which we further explored in a statistical methods of medical research paper published in 2020.
And so that I have some sources here and I thank you for your attention and welcome any questions or discussion at this point.
Thank you so much. Any questions from the room?
Okay, so yeah, I have questions.

So in the example you just showed, all the like the task becomes so, can be achievable, right?

So if the baskets, they are expected to have different benefits (indistinct), and say the 10 basket (indistinct) some other basket MEMs would allow a bigger benefit,

how will the (indistinct) scenarios?

Yeah, well, I think, if I understood your question correctly and I misheard through the phone, let me know,

but if we have different sample sizes for baskets,

which actually really corresponds

to what we’ve seen in practice for real basket trials

where they have fairly wide range of sample sizes in each basket.

I think what we would see, and let me see if I can pop back quickly to the mixed scenario results here just to illustrate some ideas.

One of the concepts here that,

so we did explicitly look at that to say like, "Well, what if one basket never gets beyond seven"

of the 25,” let’s say.
But what we can infer is that if a basket stopped early for futility, it essentially has a smaller sample size to contribute to any analysis whether or not it was a falsely stopped basket that had a 30% effect or it was truly a null basket. And so we do see in this case that the method averaging over those ideas of differential sample sizes does seem to be borrowing, appropriately depending on the context. So like the mixed scenario results here suggests limited borrowing in the presence of that uncertainty from the global scenario because we didn’t calibrate for anything else it does show more of a benefit of the stopping rate and other properties incorporating that data even in small sample sizes. And there’s also been some other work and illustrations done by Dr. Emily Zebra at the Cleveland Clinic with who I work about some of the re-analysis of oncology trials that do show even in small basket sizes, we can move that significance evaluation into a more clinically meaningful realm. <v Wayne>Thanks, so do we have other questions?</v>
Okay, so (indistinct) that’s (indistinct).

Okay, so since there are no questions let’s stop here.

(indistinct)

Yeah. Thank you all.