bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Designing fecal microbiota transplant trials that account for differences in donor stool efficacy

Scott W. Olesen[1,*], Thomas Gurry[1,2,*], Eric J. Alm[1,2,3,†]

1 Department of Biological Engineering, Massachusetts Institute of Technology, Cambridge, MA, USA

2 Center for Microbiome Informatics and Therapeutics, Massachusetts Institute of Technology, Cambridge, MA, USA

3 The Broad Institute of MIT and Harvard, Cambridge, MA, USA

***** These authors contributed equally to this work. **†** ejalm@mit.edu

Abstract

Fecal microbiota transplantation (FMT) is a highly effective intervention for patients suffering from recurrent *Clostridium difficile*, a common hospital-acquired infection. FMT's success as a therapy for *C. difficile* has inspired interest in performing clinical trials that experiment with FMT as a therapy for treating conditions like inflammatory bowel disease, obesity, diabetes, and Parkinson's disease. Results from clinical trials that use FMT to treat inflammatory bowel disease suggest that, for at least one condition beyond *C. difficile*, most FMT donors produce stool that is not efficacious. The optimal strategies for identifying and using efficacious donors have not been investigated. We therefore formulated an optimal Bayesian response-adaptive donor selection strategy and a computationally-tractable myopic heuristic. This algorithm computes the probability that a donor is efficacious by updating prior expectations about the efficacy of FMT, the placebo rate, and the fraction of donors that are efficacious. In simulations designed to mimic a recent FMT clinical trial, for which traditional power calculations predict ∼100% statistical power, we found that accounting for differences in donor efficacy reduced the predicted statistical power to ∼9%. For these simulations, using the Bayesian allocation strategy more than quadrupled the statistical power to ∼39%. We use the results of similar simulations to make recommendations about the number of patients, number of donors, and choice

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

of clinical endpoint that clinical trials should use to optimize their ability to detect if FMT is effective for treating a condition.

Author Summary

Many clinical trials test the ability of a drug to a treat a disease by comparing that drug against a placebo. In fecal microbiota transplant (FMT) trials, the "drug" is stool taken from a donor. For some diseases, however, the outcome of FMT seems to depend strongly on the choice of donor, suggesting that different donors may be producing different "drugs". Standard clinical trials are not designed to account for this possible multiplicity of drugs. We translated data from a FMT clinical trial into a model of donor stool efficacy and used this model to evaluate how different trial designs can affect a trial's ability to measure FMT's effectiveness. We found that, if only some donors produce efficacious drugs, standard clinical trial designs are likely to conclude that FMT is ineffective, thus denying patients an efficacious therapy. We also show that donor allocation strategies that adaptively use the best-performing donors can dramatically improve a trial's performance. Because FMT is not well enough understood for researchers to *a priori* identify efficacious stool, our approach makes no assumptions about the biology of the condition being treated.

Introduction

Fecal microbiota transplant (FMT), the transfer of stool from a healthy person into an ill person's gut, is a highly effective treatment for recurrent *Clostridium difficile* infections, which kill 30,000 Americans a year. Despite FMT's efficacy and increasingly widespread use, the biological mechanism by which FMT cures the infection is not fully understood [1, 2, 3, 4, 5, 6, 7]. FMT's success in treating *C. difficile* has generated interest in experiments to use FMT to treat other conditions related to the gut and the gut-associated microbiota [8, 9]. However, emerging evidence suggests using FMT for these other diseases will be more challenging. Notably, in a recent study by Moayyedi *et al.* [10] that used FMT to treat ulcerative colitis, patients appeared to respond to stool from only one of the six stool donors. Stool from all other donors was no more efficacious than placebo. These results suggest that the effectiveness of FMT can depend strongly on the choice of stool donor.

Ideally, information collected before or during a clinical trial could be used to identify which donors are efficacious. Predictions about a donor's efficacy could be repeatedly updated depending on that donor's performance, the performance of other donors, and prior expectations about donor efficacy and heterogeneity.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

There is, however, not enough clinical information about this variability or biological information about the mechanism by which FMT treats disease to create or validate a model for any particular indication. We therefore present a general model of differences in donor stool efficacy that is not specific for any clinical indication or biomarker. Using this model, we formulate an optimal, adaptive, Bayesian algorithm for allocating donors. We also formulate a computationally-tractable myopic Bayesian allocation heuristic.

Using simulations of clinical trials, we show that differences in donor efficacy and the trial's strategy for allocating donors can have substantial impacts on the trial's statistical power. We compare the performance of non-adaptive approaches to matching patients and donors against a variation of a previously-studied adaptive algorithm (a variation of the "play the winner" strategy [11, 12, 13]) and our own Bayesian heuristic. We find that, in many cases, traditional non-adaptive donor allocation strategies are likely to falsely conclude that FMT in inefficacious. Adaptive strategies, however, can substantially increase a trial's ability to detect if FMT is efficacious.

Methods

Model of differences in donor stool efficacy

The results of the trial reported in in Moayyedi *et al.* [10] raise the possibility that only some donors produce stool that is efficacious for treating ulcerative colitis. We developed a simple model of differences in the efficacy of stool produced by different donors (Fig. 1). The model we present assumes:

- Each donor produces efficacious stool or inefficacious stool. Stool from inefficacious donors is inert and has no effect beyond placebo.
- All patients receive one course of treatment, each from one donor. Multiple patients can receive stool from the same donor, but each patient receives stool from only one donor.
- Patient responses are dichotomous: a patient either responds to treatment (i.e., reaches a positive clinical endpoint) or not.
- Patients are exchangeable.

These assumptions mean that the model has only three parameters:

1. the placebo rate p_{placebo},
2. the probability p_{eff} that a patient will respond to stool from an efficacious donor, and

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

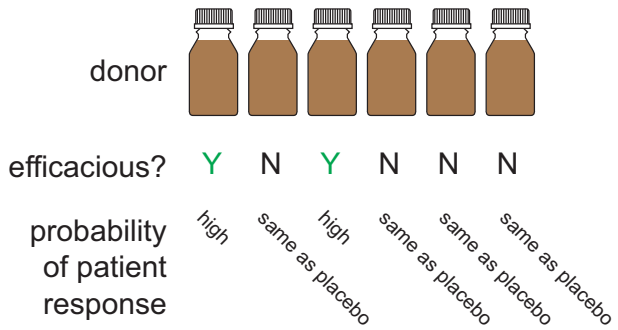

Figure 1: The model of differences in donor efficacy. In the model, donors are efficacious or not. Patients respond to FMT from an efficacious stool donor with probability p_{eff}. An FMT from an inefficacious stool donor is considered identical to a placebo, i.e., patients respond with probability p_{placebo}. The fraction of donors in the general donor population that are efficacious is f_{eff}.

Relevant trial result	Parameter	Estimated value
2 of 37 patients in placebo arm achieve remission	p_{placebo}	$2/37 = 0.054$
1 of 6 donors appeared efficacious	f_{eff}	$1/6 = 0.17$
7 of 18 patients allocated to efficacious donor achieved remission	p_{eff}	$7/18 = 0.39$

Table 1: Estimates of model parameters from the data from an ulcerative colitis clinical trial. These clinical data are drawn from the results of the trial reported in Moayyedi *et al.* [10], which used FMT to treat ulcerative colitis.

3. the frequency f_{eff} of efficacious donors among the general donor population.

These parameters can be easily related to results from clinical trials. A conversion between the clinical trial results from Moayyedi *et al.* [10] and the parameters in the model are shown in Table 1.

Further details about the model, its extensibility, and techniques for drawing random variates of its parameters are in the SI Text.

Donor allocation strategies

In simulations of clinical trials, four different strategies for choosing which donor to use with which patient were evaluated. The first two strategies, block allocation and random allocation, are non-adaptive, that is, the allocation decisions about which donor will be used for which patient can be made before the trial begins. The other two strategies, a randomized urn-based strategy and a myopic Bayesian strategy, are response-adaptive strategies, that is, the allocation about which donor to use for a patient will depend on outcomes of previous patients (Fig. 2).

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

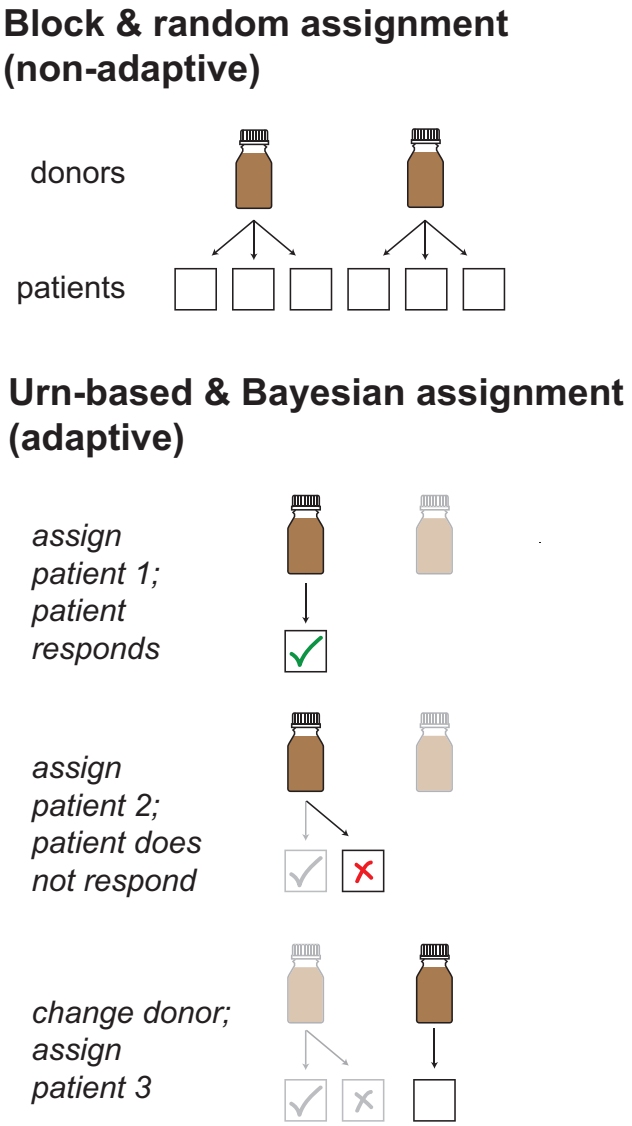

Figure 2: Adaptive donor allocations change depending on the trial's progress. In a non-adaptive allocation, the patients and donors can be matched before any patient is treated. In an adaptive allocation, the outcome of previous patients' treatments can affect the allocation of future patients. For example, if a donor is used for a patient who responds to treatment, that donor might be used again. If the patient does not respond, a different donor might be used.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Block allocation In a block allocation, patients are evenly allocated to donors. For example, if there are 30 patients and 6 donors, the first 5 patients are treated with the first donor, the second 5 patients are treated with the second donor, etc. (In clinical practice, patients would be randomized within the blocks.)

Random allocation In a random allocation, patients are allocated to donors at random. (On average, random allocations are similar to block allocations, so these two types of non-adaptive simulation yield similar results.)

Urn-based allocation Our urn-based allocation strategy is a variation on the "play-the-winner" strategies designed and studied as an ethical [14] and statistically-rigorous way to decide how to allocate patients to a treatment arm when a trial includes more than one treatment arm [15, 16, 17]. In this study, we used the generalized Pólya's urn [18] with parameters $w = 1$, $\alpha = 3$, $\beta = 0$, and without replacing the drawn ball.

Adaptive allocation strategies, like this urn-based approach, are designed to reduce the probability of a Type II error (i.e., to avoid concluding that FMT is ineffective when it actually is effective) without increasing the probability of a Type I error (i.e., without increasing the chance that a trial will conclude that FMT is effective when it is actually ineffective). If FMT is effective for some donors, this urn-based allocation should allocate more patients to efficacious donors. If FMT is ineffective (i.e., there are no efficacious donors), then an adaptive strategy should not increase the risk of a Type I error, since any and all allocation strategies will allocate only inefficacious donors.

Myopic Bayesian strategy The urn-based approach is randomized, which is desirable in clinical trials because it can reduce certain kinds of bias [11]. However, random approaches are not guaranteed to make the best choices about donor allocation. We therefore designed a myopic Bayesian deterministic donor allocation strategy that uses the results collected during a simulated trial to make the best choice about how to allocate the next patient, similar to other "bandit" problems [19]. Unless noted, the myopic Bayesian strategy was initialized with a uniform prior on the model parameters.

Computing the posterior predictive probabilities in the myopic Bayesian allocation strategy requires a computationally-intensive numerical integration. The value of the integral was computed using Monte Carlo integration with Suave (SUbregion-Adaptive VEgas), an importance sampling method combined with a globally adaptive subdivision strategy. Sampling for this integral was performed with Sobol pseudo-random numbers. The integrator was implemented in C++ using the Cuba package [20]. Wrappers for the integration routine were implemented in Python 3 and simulations were then parallelized to run on multiple cores to

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

optimize computational run time [21].

Details about the optimal Bayesian strategy and the derivation of the posterior predictive probabilities are in the SI Text.

Simulated clinical trials

The expected fraction of patients allocated to efficacious donors and the the statistical power of clinical trials using different donor allocation strategies were estimated using simulated clinical trials. In each simulation, the three model parameters (p_{placebo}, p_{eff}, f_{eff}) and the number of patients in the trial were fixed. For each combination of the trial parameters, 10,000 lists of 6 donors each were randomly generated. Donors were designated as efficacious or not efficacious by random chance according to the frequency of efficacious donors f_{eff}. The same donor lists were used for simulations for each of the allocation strategies.

In one set of simulations, the number of donors was varied among 1, 3, 5, 10, 15, and 30 donors. For those simulations, lists of 30 donors were generated for each parameter set and trial iteration. The lists were truncated for the simulations using less than 30 donors.

For each allocation strategy and donor list, a trial was simulated. In each simulation, a patient allocated to an efficacious donor responds to the treatment with probability p_{eff}. Patients allocated to inefficacious donors or to the placebo arm respond with probability p_{placebo}. For adaptive allocations, the outcomes from all the previous patients treatment were determined before the donor for the next patient was selected. An equal number of patients was allocated to the treatment and placebo arms.

Clinically-relevant parameter values Simulations were performed for all combinations of parameter values selected to reflect clinically-relevant possibilities:

- The placebo rate p_{placebo} is either 0.05 (a low placebo rate consistent with stringent, objective outcomes; e.g., endoscopic Mayo score [10, 22]) or 0.25 (a high placebo rate consistent with self-reported, subjective outcomes [23, 24]).
- The efficacy p_{eff} of efficacious donors is either 0.4 (similar to the value in Table 1) or 0.95 (efficacy of FMT to treat *C. difficile* infection).
- The frequency f_{eff} of efficacious donors is either 0.15 (similar to the value in Table 1) or 0.9 (reflective of the fact that almost any well-screened donor produces stool that can successfully treat *C. difficile* infection).

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

- The number of patients in each of the treatment and control arms is 15, 30, or 60, corresponding to a range of patient numbers typical for Phase I and small Phase II clinical trials.

Of these combinations, the set of values most similar to the one in Table 1 is $p_{\text{placebo}} = 0.05$, $p_{\text{eff}} = 0.4$, $f_{\text{eff}} = 0.15$, $N_{\text{patients}} = 30$.

Computing statistical power After determining the outcome of all the patients in the trial, the p-value of a one-sided Fisher's exact test (asserting that the response rate in the treatment arm was greater) was calculated. The proportion of simulations that produced $p < 0.05$ was the estimate of the statistical power for that allocation strategy under those trial parameters. Confidence intervals were calculated using the method of Clopper and Pearson [25]. Values are rounded to two or three significant digits.

Results

Trials using adaptive strategies allocate more patients to efficacious donors

The purpose of adaptive donor allocation strategies is to identify and use efficacious donors. We therefore expected that simulated trials that use adaptive strategies would allocate more patients to efficacious donors (compared to simulated trials that used the block or random donor allocation strategies).

For every parameter set simulated, the average fraction of patients allocated to efficacious donors was greater in the adaptive strategies (urn-based and myopic Bayesian) than in the non-adaptive strategies (block and random; Table S1). The two non-adaptive allocation strategies performed almost identically: for each parameter set, their results differed by less than 1 percentage point. The two adaptive strategies performed similarly: for half of the parameter sets, their results differed by less than 2 percentage points. In the remaining parameter sets, their results varied by between 2 and 9 percentage points.

When efficacious donors are common ($f_{\text{eff}} = 0.9$), the adaptive and non-adaptive strategies performed similarly (Fig. 3). In other cases, the performance of the two strategies differed substantially. For example, for the parameterization most similar to the one in Table 1, the random strategy allocated 15% of patients to efficacious donors while the myopic Bayesian strategy allocated 41% of patients to efficacious donors.

Trials using adaptive allocation have higher statistical power

Because trials that used the adaptive donor allocation strategies allocated more patients to efficacious donors than the trials that used the non-adaptive strategies, we expected that trials using adaptive strategies would

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

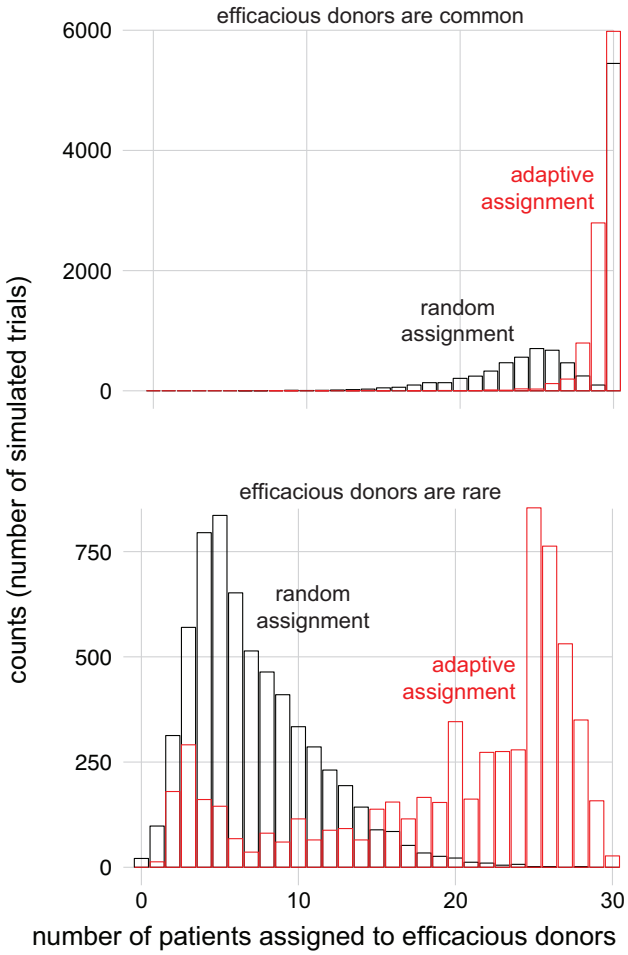

Figure 3: Adaptive strategy allocates more patients to efficacious donors. (top) For the parameterization most similar to the one in Table 1 (top), the adaptive strategy (red) allocated more patients to efficacious donors than random allocation (black) did. When efficacious donors are common (bottom; same parameters as top but with $f_{\text{eff}} = 0.9$), the two strategies allocate similar numbers of patients to efficacious donors. For visual clarity, only trials in which there was at least one efficacious donor are shown.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Trial parameters				FMT power (%)				Naive power (%)
p_{placebo}	p_{eff}	f_{eff}	N_{patients}	block	random	urn	Bayesian	
0.05	0.4	0.15	15	4.57	5.04	17.2	19.1	68.1
			30	8.44	8.89	33.7	39.4	93.8
			60	23	24.1	53.5	58	100
		0.9	15	59.1	59.7	63.4	64.3	67.8
			30	87.7	87.4	92.1	93.2	94.2
			60	99.2	99.1	99.8	99.9	100
	0.95	0.15	15	19.9	21.3	60.3	60.7	100
			30	33.9	34.2	61.4	61.6	100
			60	55.9	53.2	63.2	63.1	100
		0.9	15	99.8	99.5	100	100	100
			30	100	100	100	100	100
			60	100	100	100	100	100
0.25	0.4	0.15	15	2.8	2.88	3.43	3.77	11.4
			30	4.44	4.55	7.16	7.27	25.5
			60	6.55	6.15	13.6	13.9	47.3
		0.9	15	10.5	10.3	10.5	10.7	11.7
			30	21.6	21.3	23	23	25
			60	40.7	40.7	44.5	44.8	47.2
	0.95	0.15	15	9.04	10.3	38.6	47.4	99.4
			30	20	20.3	60	60.1	100
			60	32.6	32.7	63.4	62.7	100
		0.9	15	94.3	93.9	98.2	99	99.6
			30	99.6	99.4	100	100	100
			60	100	100	100	100	100

Table 2: Adaptive strategies yield clinical trials with higher statistical power. "FMT power" is the power computed by simulating the results of trials that would occur if the frequency of efficacious donors if f_{eff}. "Naive power" is the power computed in the situation in which all donors are efficacious (i.e., $f_{\text{eff}} = 1.0$). All 95% confidence intervals on these values are within 1 percentage point of the reported value and are not shown.

have greater statistical power.

The adaptive strategies consistently yielded higher statistical powers than the non-adaptive strategies (Table 2). When efficacious donors are rare, the performance gap is larger. For example, for the parameterization most similar to the one in Table 1, a trial that uses random allocation is expected to have 9% power, while the myopic Bayesian strategy can deliver 39% power. The gap in performance is smallest when selecting a donor at random is likely to yield an efficacious donor: among the trials with $f_{\text{eff}} = 0.9$, the adaptive and non-adaptive statistical powers differed by less than 6 percentage points.

Low statistical powers when f_{eff} is small are likely due to the fact that if *all* the available donors are not efficacious, then no allocation strategy should make a trial achieve significance. For example, if only 15% of donors are efficacious ($f_{\text{eff}} = 0.15$) and there are only 6 donors (the number used in these simulations),

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

then we expect that 38% of trials will have no good donors (using the binomial distribution function). We therefore separately analyzed the simulated trials in which no donors were efficacious and the simulated trials in which at least one donor was efficacious (Table S2). When no donors are efficacious, trials with adaptive or non-adaptive strategies have ∼0% power. Among trials with at least one efficacious donor, the difference in statistical power between adaptive and non-adaptive strategies is greater than the difference computed using the results of all trials.

Conversely, the power computed in traditional calculations that do not account for differences in donor efficacy (i.e., that assume that all donors are efficacious, or equivalently $f_{\text{eff}} = 1.0$) is, in many cases, substantially higher than the power computed when accounting for differences in donor efficacy (Table 2, column "Naive powers"). For example, for the parameter set most similar to the one in Table 1, the naive calculation predicts 94% power, but the calculation that accounts for differences in donor efficacy predicts only 9% power for non-adaptive allocation strategies. The differences between the powers computed by the naive method and our approach is largest when f_{eff} is small.

Performance of adaptive strategies depend on their parameterization

In these simulations, we varied the actual values of p_{placebo} and p_{eff} but we always initialized the adaptive algorithms the same ways. To determine the sensitivity of the adaptive allocation algorithms' performance to their initialization, we simulated trials in which the actual model parameters were fixed but the algorithms' initializations varied (Table S3 and Table S4). The myopic Bayesian algorithm's performance was mostly robust to the parameterization of its prior distribution except when the prior was strong and inaccurate. Accurate priors, weak priors, and uniform priors provide comparable performance. In contrast, the urn algorithm delivered widely varying powers, from 15% to 40%, depending on its parameterization.

Increasing the number of available donors benefits the adaptive Bayesian strategy

Increasing the number of available donors increases the probability that at least one of them will be efficacious. We therefore determined, for each donor allocation strategy, the number of available donors that optimized the trial's expected power. Simulations showed that increasing the size of the donor "pool" almost always increased the power of trials using the myopic Bayesian donor allocation but, depending on the parameter set, could increase or decrease the power of trials using other allocation strategies (Table S5).

Table 3 shows how donor selection strategy and model parameter values affects the optimal number of donors. Notably, when efficacious donors are uncommon ($f_{\text{eff}} = 0.15$) and only moderately efficacious

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Trial parameters			Optimal N_{donors}			
f_{eff}	p_{eff}	p_{placebo}	random	block	urn	Bayesian
0.15	0.4	0.05	1	1	10	10–30
0.15	0.4	0.25	1	1	5	10
0.15	0.95	0.05	3–15	3–5	15–30	15–30
0.15	0.95	0.25	3	3	10	10–30
0.9	0.4	0.05	1–30	3–30	3–30	3–30
0.9	0.4	0.25	1–30	1–30	1–30	5–30
0.9	0.95	0.05	3–30	3–30	3–30	3–30
0.9	0.95	0.25	3–30	3–30	3–30	3–30

Table 3: The optimal number of donors varies by donor selection strategy and model parameter values. For each parameter value, trials were simulated using 1, 3, 5, 10, 15, and 30 donors. (The number of patients was fixed at 30.) The number of donors that optimized the expected power was identified. If multiple numbers of donors yielded powers within 0.05 of the optimal value, all those numbers are reported as a range.

($p_{\text{eff}} = 0.4$), the non-adaptive strategies perform optimally when only one donor is used. In other words, when using non-adaptive strategies in this parameter regime, it is wiser to take the 15% chance of picking a single efficacious donor than it is to distribute patients across many donors, allotting around 15% of them to efficacious donors. In contrast, the myopic Bayesian donor allocation almost always benefits from a larger donor pool.

Discussion

Model limitations

We chose to use a simple model for a simple use case because, in the absence of data about the treatment histories of hundreds of patients using dozens of donors, we do not believe that more complicated models will be more useful to aiding trial design. The model's greatest weakness is that it cannot be validated, but it is exactly the model's purpose to improve the probability of collecting the kind of information that could validate or invalidate it. In light of the dearth of data, we developed a simple model, and it could be that the simplifications we made limit the model's validity. For example, the model assumes that each donor produces efficacious stool or inefficacious stool (when in fact there is probably day-to-day and donor-to-donor variation in stool efficacy) and that all patients receive one course of treatment (while, say, patients who do not respond to a first treatment might be treated with stool from a different donor).

In our simulations, we assumed that the outcome from all the previous patients treatments are known before the donor for the next patient is selected. In reality, patients in an FMT trial overlap. The urn-based

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

method can still be used for overlapping patients [18], but the myopic Bayesian method would require some modification. For example, clinicians could choose to consult the myopic Bayesian's rankings of donors intermittently, or patients could be allocated in proportion to the predicted probabilities of successful treatments.

Differences in donor efficacy should be accounted for in trial design

Our results entail recommendations to clinicians. First, the powers we computed here are, in many cases, well below the powers computed assuming that all donors are efficacious. We therefore encourage researchers to consult our predictions about statistical power when deciding on the size of their trials.

Second, a high placebo rate can substantially decrease the statistical power of an FMT trial. We therefore encourage researchers to use the most stringent outcome measurement possible (e.g., an endoscopic Mayo score for inflammatory bowel disease).

Third, adaptive donor allocation strategies consistently delivered higher statistical power than traditional, non-adaptive approaches. We therefore recommend that researchers use such an adaptive strategy. The urn-based strategy has the advantages that similar response-adaptive strategies may be familiar to clinicians, it is randomized, and it is simple to implement. However, an urn-based strategy needs to be carefully parameterized: a badly-parameterized urn-based strategy performs similarly to random allocation. The adaptive Bayesian donor allocation algorithm performs well even when using the "default" settings (a uniform prior) but is complex and deterministic. To fully leverage this strategy, a clinician would need to consult the algorithm's output after every patient outcome and follow the algorithm's deterministic instructions, which might introduce bias.

Fourth, adaptive algorithms benefit from having access to a "bank" of 10 or more donors. Researchers hoping to achieve the full benefits of adaptive donor selection must be prepared to change donors multiple times during the trial.

Finally, researchers reporting about FMT trials should include information about the donors, notably how many donors were used and what proportion of patients alloted to each donor responded to treatment. This information will help future researchers account for differences in donor efficacy.

Future research may identify mechanistic explanations of FMT's efficacy

The adaptive allocation strategies we described here have a narrow aim: to increase the number of successful patient outcomes in a trial. In theory, an adaptive trial design is capable of more. For example, if it were hypothesized that FMT succeeded or failed because of the presence or absence of some particular microbial

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

species in the donor's stool, then an adaptive trial design could recommend donor choices that aim to identify that critical species.

We did not pursue a hypothesis-centric approach because we believe it is premature. Even the mechanism by which FMT treats *C. difficile*, the most well-studied case, remains unclear. We expect that strong hypotheses about mechanism will come from retroactive comparison of efficacious vs. inefficacious stool *after* clinical trials have definitively show that FMT is effective for treating some disease. Our study aims to do exactly this. Until then, we hope that our results about adaptive donor allocation help more patients benefit from FMT and will help clinicians identify those conditions that FMT can treat.

Acknowledgments

This material is based on work supported by the National Science Foundation Graduate Research Fellowship under Grant No. 1122374 and by the Center for Microbiome Informatics and Therapeutics at MIT.

References

[1] Cohen SH, Gerding DN, Johnson S, Kelly CP, Loo VG, McDonald LC, et al. Clinical Practice Guidelines for *Clostridium difficile* Infection in Adults. Infect Control Hosp Epidemiol. 2010;31(5):431–455. doi:10.1086/651706.

[2] van Nood E, Vrieze A, Nieuwdorp M, Fuentes S, Zoetendal EG, de Vos WM, et al. Duodenal Infusion of Donor Feces for Recurrent Clostridium difficile. N Engl J Med. 2013;368(5):407–415. doi:10.1056/NEJMoa1205037.

[3] Kelly CP, LaMont JT. *Clostridium difficile* – More Difficult Than Ever. N Engl J Med. 2008;359(18):1932–1940. doi:10.1056/NEJMra0707500.

[4] Alang N, Kelly CR. Weight Gain After Fecal Microbiota Transplantation. Open Forum Infect Dis. 2015;2(1). doi:10.1093/ofid/ofv004.

[5] Li SS, Zhu A, Benes V, Costea PI, Hercog R, Hildebrand F, et al. Durable coexistence of donor and recipient strains after fecal microbiota transplantation. Science. 2016;352(6285):586–589. doi:10.1126/science.aad8852.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[6] Kassam Z, Lee CH, Yuan Y, Hunt RH. Fecal Microbiota Transplantation for *Clostridium difficile* Infection: Systematic Review and Meta-Analysis. Am J Gastroenterol. 2013;108(4):500–508. doi:10.1038/ajg.2013.59.

[7] Khoruts A, Sadowsky MJ. Understanding the mechanisms of faecal microbiota transplantation. Nat Rev Gastroenterol Hepatol. 2016;doi:10.1038/nrgastro.2016.98.

[8] Smits LP, Bouter KEC, de Vos WM, Borody TJ, Nieuwdorp M. Therapeutic Potential of Fecal Microbiota Transplantation. Gastroenterology. 2013;145(5):946–953. doi:10.1053/j.gastro.2013.08.058.

[9] Sadowsky MJ, Khoruts A. Faecal microbiota transplantation is promising but not a panacea. Nat Microbiol. 2016;1. doi:10.1038/nmicrobiol.2016.15.

[10] Moayyedi P, Surette MG, Kim PT, Libertucci J, Wolfe M, Onischi C, et al. Fecal Microbiota Transplantation Induces Remission in Patients With Active Ulcerative Colitis in a Randomized Controlled Trial. Gastroenterology. 2015;149(1):102–109.e6. doi:10.1053/j.gastro.2015.04.001.

[11] Cook TD, DeMets DL. Introduction to Statistical Methods for Clinical Trials. Chapman & Hall/CRC; 2008.

[12] Bartlett RH, Roloff DW, Cornell RG, Andrews AF, Dillon PW, Zwischenberger JB. Extracorporeal Circulation in Neonatal Respiratory Failure: A Prospective Randomized Study. Pediatrics. 1985;76(4):479–487.

[13] Chow SC, Chang M. Adaptive Methods in Clinical Trials. Chapman & Hall/CRC; 2007.

[14] Hung JHM, O'Neill RT, Wang SJ, Lawrence J. A Regulatory View of Adaptive/Flexible Clinical Trial Design. Biom J. 2006;48(4):565–573. doi:10.1002/bimj.200610229.

[15] Wei LJ, Smythe RT, Lin DY, Park TS. Statistical Inference with Data-Dependent Treatment Allocation Rules. J Am Stat Assoc. 1990;85(409):156–162. doi:10.1080/01621459.1990.10475319.

[16] Wei LJ, Durham S. The Randomized Play-the-Winner Rule in Medical Trials. J Am Stat Assoc. 1978;73(364):840–843. doi:10.1080/01621459.1978.10480109.

[17] Zelen M. Play the Winner Rule and the Controlled Clinical Trial. J Am Stat Assoc. 1969;64(325):131–146. doi:10.1080/01621459.1969.10500959.

[18] Wei LJ. The Generalized Polya's Urn Design for Sequential Medical Trials. Ann Stat. 1979;7(2):291–296.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

[19] Berry DA, Fristedt B. Bandit Problems: Sequential Allocation of Experiments. Springer Netherlands; 1985.

[20] Hahn T. Cuba – a library for multidimensional numerical integration. Comput Phys Commun. 2005;168(2):78–95. doi:10.1016/j.cpc.2005.01.010.

[21] Tange O. GNU Parallel – The Command-Line Power Tool. ;login: The USENIX Magazine. 2011;36(1):42–47. doi:10.5281/zenodo.16303.

[22] Rossen NG, Fuentes S, van der Spek MJ, Tijssen JG, Hartman JHA, Duflou A, et al. Findings From a Randomized Controlled Trial of Fecal Transplantation for Patients With Ulcerative Colitis. Gastroenterology. 2015;149(1):110–118. doi:10.1053/j.gastro.2015.03.045.

[23] Su C, Lichtenstein GR, Krok K, Brensinger CM, Lewis JD. A meta-analysis of the placebo rates of remission and response in clinical trials of active Crohn's disease. Gastroenterology. 2004;126(5):1257–1269. doi:10.1053/j.gastro.2004.01.024.

[24] Su C, Lewis JD, Goldberg B, Brensinger C, Lichtenstein GR. A Meta-Analysis of the Placebo Rates of Remission and Response in Clinical Trials of Active Ulcerative Colitis. Gastroenterology. 2007;132(2):516–526. doi:10.1053/j.gastro.2006.12.037.

[25] Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. Biometrika. 1934;26(4):404–413. doi:10.1093/biomet/26.4.404.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Supplementary Tables

p_{placebo}	p_{eff}	f_{eff}	N_{patients}	block		random		urn		Bayesian	
				mean	s.d.	mean	s.d.	mean	s.d.	mean	s.d.
0.05	0.4	0.15	15	0.15	0.15	0.15	0.17	0.29	0.30	0.32	0.33
			30	0.15	0.14	0.15	0.16	0.36	0.35	0.41	0.38
			60	0.15	0.14	0.15	0.15	0.44	0.39	0.49	0.42
		0.9	15	0.9	0.12	0.9	0.14	0.95	0.075	0.96	0.068
			30	0.9	0.12	0.9	0.13	0.97	0.051	0.98	0.041
			60	0.9	0.12	0.9	0.13	0.98	0.03	0.99	0.022
	0.95	0.15	15	0.15	0.15	0.15	0.17	0.44	0.36	0.51	0.42
			30	0.15	0.15	0.15	0.16	0.52	0.42	0.56	0.45
			60	0.15	0.15	0.15	0.15	0.57	0.45	0.6	0.47
		0.9	15	0.9	0.13	0.9	0.14	0.97	0.047	0.99	0.028
			30	0.9	0.12	0.9	0.13	0.98	0.027	1.0	0.013
			60	0.9	0.12	0.9	0.13	0.99	0.014	1.0	0.0069
0.25	0.4	0.15	15	0.15	0.15	0.15	0.17	0.20	0.25	0.21	0.29
			30	0.15	0.15	0.15	0.16	0.24	0.30	0.24	0.34
			60	0.15	0.15	0.15	0.15	0.29	0.34	0.28	0.38
		0.9	15	0.9	0.13	0.9	0.14	0.93	0.14	0.94	0.15
			30	0.9	0.12	0.9	0.13	0.94	0.13	0.95	0.14
			60	0.9	0.12	0.9	0.13	0.96	0.11	0.96	0.14
	0.95	0.15	15	0.15	0.15	0.15	0.17	0.36	0.34	0.45	0.42
			30	0.15	0.15	0.15	0.16	0.45	0.38	0.52	0.44
			60	0.15	0.15	0.15	0.15	0.53	0.42	0.57	0.46
		0.9	15	0.9	0.13	0.9	0.14	0.96	0.076	0.99	0.056
			30	0.9	0.12	0.9	0.13	0.98	0.048	0.99	0.029
			60	0.9	0.12	0.9	0.13	0.99	0.029	1.0	0.016

Table S1: Fraction of patients allocated to efficacious donors. The same simulated trials were analyzed to create Table 2 and this table.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Parameters		Simulation conditions		Power (%)			
p_{eff}	f_{eff}	Donor pool quality	$N_{\text{simulations}}$	random	block	urn	Bayesian
0.4	0.15	all trials	10000	8.9	8.4	34	39
0.4	0.15	no efficacious donors	3794	0.37	0.29	0.4	0.16
0.4	0.15	some efficacious donors	6206	14	13	54	63
0.4	0.9	some efficacious donors	10000	87	88	92	93
0.95	0.15	all trials	10000	34	34	61	62
0.95	0.15	no efficacious donors	3851	0.62	0.44	0.34	0.31
0.95	0.15	some efficacious donors	6149	55	55	100	100
0.95	0.9	some efficacious donors	10000	100	100	100	100

Table S2: Power of simulated clinical trials, conditioned on presence of efficacious donors. A subset of the data used to create Table 2 (only those simulations with $N_{\text{patients}} = 30$ and $p_{\text{placebo}} = 0.05$) was analyzed by separately estimating the power for the trials in which no donors were efficacious ("no efficacious donors") and in which at least one donor was efficacious ("some efficacious donors"). "All simulations" shows the same data as in Table 2. For $f_{\text{eff}} = 0.9$, none of the 10,000 simulations had a donor pool with no efficacious donors. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

	Hyperparameter values						
Hyperparameterization	A_{placebo}	B_{placebo}	A_{peff}	B_{peff}	A_{feff}	B_{feff}	Power (%)
From Table 1	2	35	7	11	1	5	41.1
Weak, accurate	0.2	3.5	0.7	1.1	0.1	0.5	41.2
Strong, accurate	20	350	70	110	10	50	40.4
Weak, inaccurate	0.4	3.3	1.4	0.4	0.2	0.4	40.9
Strong, inaccurate	40	330	140	40	20	40	34.4
Uniform	n.a.	n.a.	n.a.	n.a.	n.a.	n.a.	39.6

Table S3: Sensitivity of simulated clinical trial power to parameterization of the myopic Bayesian algorithm parameters. Starting from the parameter described in Table 1, 10,000 trials using the myopic Bayesian donor allocation were simulated for each of six different parameterizations of the Bayesian algorithm. "Accurate" means that the prior is centered around the true value (i.e., that $\frac{A}{A+B}$ equals the true value); "inaccurate" means that the prior is centered at approximately double the true value (but that $A + B$ has been held constant). "Strong" means that every hyperparameter is ten-fold greater; "weak" means that every hyperparameter is ten-fold smaller. "Uniform" means a uniform prior was used for all parameters. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

w	α	β	replace?	Power (%)
1	3	0	true	28.5
			false	35.6
		1	true	13.9
			false	14.2
	6	0	true	33.2
			false	41
		1	true	16.2
			false	16.4
3	3	0	true	19.7
			false	29.1
		1	true	13.2
			false	13.1
	6	0	true	24.8
			false	32.6
		1	true	15.6
			false	15.3

Table S4: Sensitivity of simulated clinical trial power to urn parameterization. Using the parameter set described in Table 1, 10,000 trials using the urn-based donor allocation were simulated for each of several combinations of the parameters for the urn model as described in [18]. We also vary whether drawn balls are replaced or not. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.

bioRxiv preprint doi: https://doi.org/10.1101/065383; this version posted July 25, 2016. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY-ND 4.0 International license.

Parameters				Power (%)			
f_{eff}	p_{eff}	p_{placebo}	N_{donors}	random	block	urn	Bayesian
0.15	0.4	0.05	1	15.5	15.3	15.5	14.8
			3	12.4	12.2	28.4	30.9
			5	9.87	9.67	33.8	38
			10	7.49	6.93	38.2	44.3
			15	6.91	6.38	34.6	45.9
			30	6.17	5.26	16.6	45.9
		0.25	1	6.08	6.15	5.89	5.58
			3	4.64	4.21	6.7	6.89
			5	4.65	4.7	7.28	7.1
			10	4.37	4.18	6.2	7.52
			15	4.48	4.09	5.41	7.07
			30	4.25	4.3	4.67	6.89
	0.95	0.05	1	15.2	15.1	15	14.9
			3	34.4	37.2	39	39.2
			5	35	36.1	56.3	56.5
			10	33.7	33.3	80.1	80.4
			15	33.5	32.6	91	90.6
			30	31.6	30.4	87.4	93.9
		0.25	1	17.1	17.3	17	16.8
			3	23.1	22.6	39	39.9
			5	21.3	20.9	55.2	55.5
			10	18.8	17.7	69.7	69.7
			15	17.4	16.8	65.4	69.9
			30	16.4	15.4	43.2	69.4
0.9	0.4	0.05	1	84.9	84.7	84.6	84.3
			3	85.9	85.8	91.8	92.4
			5	87.4	87.2	92.4	92.4
			10	88.4	88.9	92.3	92.8
			15	88	89	91.8	92.9
			30	89.1	89.3	91.1	92.9
		0.25	1	22	22.3	22.5	22.1
			3	21.8	21.2	23.1	22.7
			5	21.7	21.3	23.1	23.2
			10	21.7	21.3	22.7	23.7
			15	22.2	21.5	22.5	24
			30	21.9	21.6	22.1	23.5
	0.95	0.05	1	90.1	90.1	90.1	90.4
			3	99.5	99.7	99.9	99.9
			5	100	100	100	100
			10	100	100	100	100
			15	100	100	100	100
			30	100	100	100	100
		0.25	1	90.4	90.4	90.4	90.5
			3	97.7	98.1	99.9	99.9
			5	99.1	99.4	100	100
			10	99.8	99.9	100	100
			15	99.8	99.9	100	100
			30	99.9	100	100	100

Table S5: Dependence of simulated clinical trial power on number of available donors. For each parameter value, trials were simulated using 1, 3, 5, 10, 15, and 30 donors. The number of patients was fixed at 30. All 95% confidence intervals are within 1 percentage point of the reported value and are not shown.