SUPPLEMENT

Section S1. Simulation settings and trial design

Study design: We considered fixed-randomized and outcome-adaptive randomized multi-arm trial designs. For both designs, we tested the null hypothesis of no treatment effect against the alternative of a positive treatment effect, computing at the end of the trial for each experimental arm a two-sample Z-statistics for proportions. Experimental arms with p-value>0.05 were classified as ineffective. For the fixed randomization we followed the original plan provided in the MOTIVATE and REMIND papers.[1-3]

For the outcome-adaptive design, allocation probabilities were updated M times during the trial at equal (enrollment) intervals, following a Bayesian adaptive randomization (BAR) allocation rule.[4] Specifically, the allocation probability to effective arms is proportional to $pr(\pi_a > \pi_0 | \, Accrued\ Data)^{h(i)}$, where π_a is the probability of a positive outcome in one of the effective arms, to which we assign uniform priors. The function $h(i) = N(i)/N$, with $N(i)$ being the number of patients currently enrolled in the trial, modulates the tradeoff of exploration-exploitation. Patients assigned to usual care were allocated with probability proportional to: $exp\{\max_{a>0} N_a - N_0\}/A$, where $N_a$is the number of accrued subjects in arm $a = 0, \dots, A$. We also considered J (J=1 or 2) futility interim analyses (IAs). Experimental arms were dropped early for futility when the posterior probability $pr(\pi_a > \pi_0 | Accrued\ Data)$ was less than a pre-specified threshold. At each IA, the sample size was recalculated: If one or more arms were dropped, sample size was adjusted according to the following formula: $N_{adj} = \mathrm{N} - \frac{N}{A+1} D + \sum_{c \in I_D} N_c$, where D

indicates the number of dropped arms, I_D the indices of the arm that are dropped and N_c the number of patients already enrolled in arm c.

As in the original trials, we chose not to formally adjust for multiple testing because Bonferroni corrections were deemed too conservative and reasoning that if each exposure was compared with control in a separate trial, no adjustment would be necessary.[5,6]

Simulation scenarios: We considered three simulation scenarios based on these two completed trials. In all scenarios, a maximum of *N* subjects are allocated to the control arm or *A* experimental arms, where *N* matches the sample size of the original trial. We used 3 simulation scenarios (**Table S1**) for each original trial and computed the operating characteristics generating 10,000 *in silico* trials for each scenario. We assumed that interim analyses would each take 30 days, and the average duration of the trials was calculated as a combination of these 30-day interim analyses and lengths of time to measure outcomes across the 10,000 replications. We used Julia programming software (version 7.1) to conduct the simulations, and the R software (version 4.1.2) to produce the figures.

In Scenario 1, we assumed that there were different treatment effects across the experimental arms compared with control. One experimental arm was modestly effective, one was highly effective, and one was ineffective. In Scenario 2, only 1 experimental intervention arm effective (which we hypothesized would favor an adaptive RCT). In Scenario 3, all experimental arms were equally effective (which we hypothesized would favor a fixed-randomization RCT).

For each trial, we then used the same effectiveness assumptions by arm as Scenarios 1-3 but instead we considered shorter windows to measure the primary outcome so that

the total duration of the adaptive trials, including follow-up and interim analyses, matched the original, published trials. In these scenarios, we assumed that only 1 interim analysis was conducted, as that would be more practical in the shorter duration.

Specifically, for MOTIVATE, we used a 1-month window to measure the primary outcome of influenza vaccination receipt rather than 4 months. Similarly, for REMIND, we shortened the measurement window for the primary outcome of optimal adherence to 3 months rather than 12 months. We chose these time periods based on prior literature that demonstrates that patient letters typically lead to action over the short term[1] and because short-term adherence measures align well with longer-term measures of adherence, particularly for prevalent user of medications.[7,8]

Additional MOTIVATE assumptions

For the 5-arm MOTIVATE trial, we compared the (response) adaptive randomizations with the fixed allocation ratios 10:2:2:3:3 of the 228,000 participants as in the original trial and modeled 3 scenarios.

In Scenario 1, we modeled that Arm 2 (letter from National Vaccine Program Office) achieved a 10% effect size, Arms 3 (letter from US Surgeon General) and 4 (letter from US Surgeon General and implementation intention prompt) achieved a 5% effect size, and Arm 5 (letter from US Surgeon General and active choice implementation intention prompt) was no different than control.

In Scenario 2, we simulated that Arms 2 achieved a 10% effect size, and Arms 3-5 were no different than control.

In Scenario 3, we modeled a 5% effect size each for Arms 2-5 compared with control.

These assumptions were made because the original trial assumed the ability to detect a 5% difference in vaccination rates between arms, and these scenarios illustrate how sample size requirements change based upon the magnitude of differences between the arms and usual care and between the arms themselves, such as for example, one arm being much more effective than the others or several arms being much more effective than usual care. The probabilities of positive outcomes are shown in **Table S1.** We set the pre-specified threshold for the posterior probability to say that an arm is ineffective as 0.9, as has been done in prior work.

For MOTIVATE Scenarios 1-3, we set M=6 (i.e., 6 blocks of time of 4 months each), and we performed 2 interim analyses in the simulation after blocks 2 and 4, respectively. This means that subjects are enrolled, they finished follow-up, an interim analysis is conducted, and then the study proceeds to the next block. We used 4-month follow-up in these scenarios, as the range for follow-up for influenza vaccination outcomes in trials ranges from 1-6 months typically, and 4 months was used in the original trial. Using these assumptions, the trial can last up to 24 months for Scenarios 1-3. While this may not be practical in reality, we use this to show the importance of relatively rapid observable outcomes for adaptation in trials. Then, we set M=4 (i.e., 4 blocks of time of 1 month each) and performed 1 interim analysis in the simulation after the 2nd block (Table 2 in manuscript text). We assumed a similar rate of effectiveness, as the impact of patient communication for vaccination outreach for influenza typically is relatively rapid.

Additionally, to further evaluate potential benefits of adaptive randomization, we conducted one more simulation scenario based on MOTIVATE. In specific, we reduced the sample of the original trial to 6,300 subjects, corresponding with the original 10:2:2:3:3 allocation ratio, to an average number of patients of (3,150, 630, 630, 945, 945) for the control arm and the four intervention arms, respectively. With this sample size, we would have 80% power to the detect an effect on Arm 1 when it has a probability of observing a positive outcome of 0.7, and a power of 90% when Arm 1 has a probability of observing a positive outcome of 0.75.

Then, we tuned a Bayesian adaptive design to have the same average sample size of the fixed randomization when all the four experimental arms share the same probability of observing a positive outcome of 0.7 (Scenario 3). We performed one interim analysis when half of the patients have been enrolled and updated the probabilities to be enrolled to each of the arms four times during the trial. The maximum number of patients that can be enrolled in the adaptive design is 6,600. The results are presented in **Table S2**. In Scenario 1 and 2, the adaptive randomization increases power while reducing the average sample size. For example, in Scenario 1 and 2 the power of detecting an effect in arm 1 increases to 0.99 compared 0.90 of the fixed randomization, while the average sample size is of 5,231, and 3,985 compared to 6,300 of the fixed randomization. In Scenario 3 the adaptive randomization has the same average sample size than the fixed randomization. The slightly higher power in Table S2 is given by the fact that in some replicates of the trial the overall sample size is higher than the fixed randomization. It is worth noticing that the probability that a non-effective arm is removed during the interim analysis is approximately 60%, while arms with a probability of positive outcome of 0.70 are flagged as ineffective less

than 5% of the time. The arms with a probability of positive outcomes of 0.75 are never excluded from the trial.

Additional REMIND assumptions

For the 4-arm REMIND trial, we focused on "Block A – Chronic Diseases" for the simulation and assumed a 1:2:2:2 allocation ratio of 22,163 participants as in the trial. As before, we otherwise used the same assumptions as in the trial and conducted 3 simulations with differences in the probability of positive outcomes (shown in **Table S1**). As in MOTIVATE, we set the pre-specified threshold for the posterior probability to say that an arm is ineffective as 0.9, as has been done in prior work.

For REMIND, in Scenarios 1, we modeled that Arms 2, 3, and 4 achieved effect sizes of 8%, 5%, and no difference versus the control arm, respectively. In Scenarios 2, we modeled that Arm 2 achieved an effect size of 8% and Arms 3 and 4 were no different than control. In Scenarios 3, we modeled that Arms 2, 3 and 4 each achieved effect sizes of 5% versus control. These assumptions were based on the original trial sample size calculations and to demonstrate how sample size requirements would change based on the magnitude of difference across the arms (i.e., one arm being much more effective than other arms, as in Scenario 2). For Scenarios 1-3, we set M=5 (i.e., 5 blocks of 12 months of time based on the 12-month adherence outcome), and we performed 2 interim analyses in the simulation after blocks 2 and 4. We then considered a shorter length trial, setting M=4 (i.e., 4 blocks of 3 months of time based on a 3-month adherence outcome), and we performed 1 interim analysis in the simulation after block 2. Of note, if using a 12-month outcome with the same assumptions, this scheme would expect to collect data in maximum 4 years (as seen in

Scenarios 1-3). We set the pre-specific threshold for the posterior probability to say that an arm is ineffective to 0.5.

Figure S1: Case example of NUDGE-EHR adaptive trial

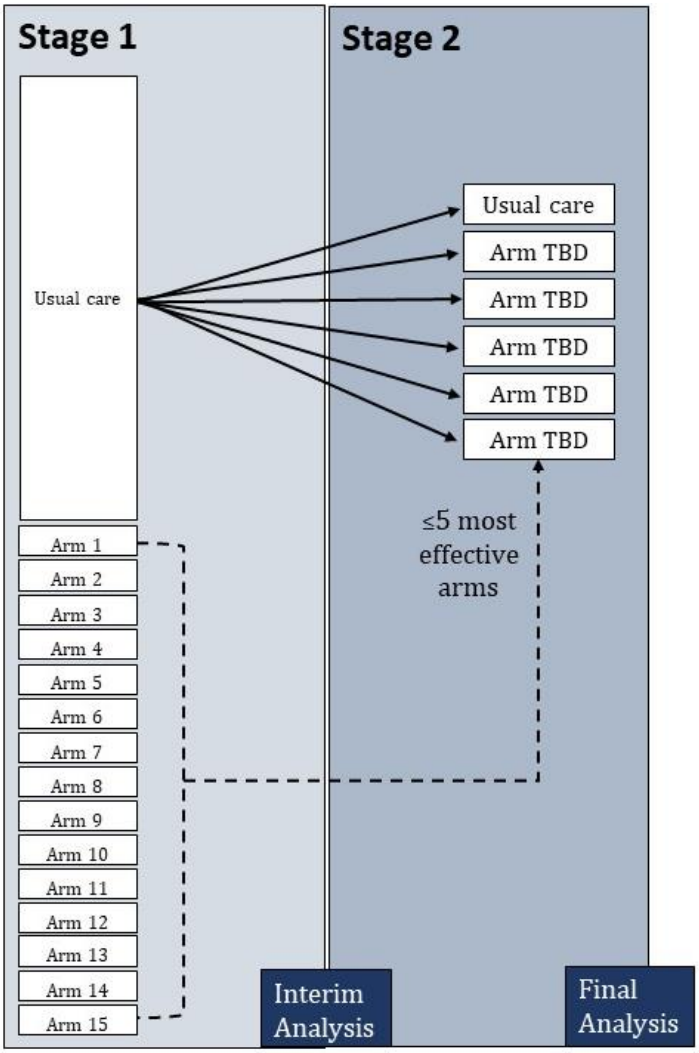

Table S1. Precision of the estimated response rates for the REMIND and MOTIVATE trials.

Scenario	Average probability of observing a positive outcome and standard deviation per arm				
	Control	Arm 1	Arm 2	Arm 3	Arm 4
MOTIVATE trial					
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)	0.650 (0.001)	0.750 (0.003)	0.700 (0.003)	0.700 (0.002)	0.650 (0.003)
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)	0.650 (0.001)	0.750 (0.003)	0.650 (0.003)	0.650 (0.003)	0.650 (0.003)
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)	0.650 (0.001)	0.700 (0.003)	0.700 (0.003)	0.700 (0.002)	0.700 (0.002)
Adaptive: *Same outcome measurement window as original trial and two 30-day interim analyses*					
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)	0.650 (0.002)	0.750 (0.003)	0.700 (0.003)	0.700 (0.003)	0.650 (0.005)
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)	0.650 (0.003)	0.750 (0.005)	0.650 (0.005)	0.650 (0.005)	0.650 (0.005)
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)	0.650 (0.002)	0.700 (0.002)	0.700 (0.002)	0.700 (0.002)	0.700 (0.002)
Adaptive: *Shorter outcome measurement window and one 30-day interim analysis*					
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)	0.650 (0.002)	0.750 (0.003)	0.700 (0.003)	0.700 (0.003)	0.650 (0.004)
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)	0.650 (0.002)	0.750 (0.004)	0.650 (0.004)	0.650 (0.004)	0.650 (0.004)
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)	0.650 (0.002)	0.700 (0.002)	0.700 (0.002)	0.700 (0.002)	0.700 (0.002)
REMIND trial					
Scenario 1: (0.02, 0.10, 0.07, 0.02)	0.020 (0.002)	0.100 (0.004)	0.070 (0.003)	0.020 (0.002)	
Scenario 2: (0.02, 0.10, 0.02, 0.02)	0.020 (0.002)	0.100 (0.003)	0.020 (0.002)	0.020 (0.002)	
Scenario 3: (0.02, 0.07, 0.07, 0.07)	0.020 (0.002)	0.070 (0.003)	0.070 (0.003)	0.070 (0.003)	
Adaptive: *Same outcome measurement window as original trial and two 30-day interim analyses*					
Scenario 1: (0.02, 0.10, 0.07, 0.02)	0.020 (0.002)	0.100 (0.008)	0.070 (0.006)	0.020 (0.004)	
Scenario 2: (0.02, 0.10, 0.02, 0.02)	0.020 (0.002)	0.100 (0.009)	0.020 (0.004)	0.020 (0.004)	
Scenario 3: (0.02, 0.07, 0.07, 0.07)	0.020 (0.002)	0.070 (0.007)	0.070 (0.007)	0.070 (0.007)	
Adaptive: *Shorter outcome measurement window and one 30-day interim analysis*					
Scenario 1: (0.02, 0.10, 0.07, 0.02)	0.020 (0.002)	0.100 (0.007)	0.070 (0.006)	0.020 (0.003)	
Scenario 2:	0.020	0.100	0.020	0.020	

(0.02, 0.10, 0.02, 0.02)	(0.002)	(0.007)	(0.003)	(0.003)	
Scenario 3: (0.02, 0.07, 0.07, 0.07)	0.020 (0.002)	0.070 (0.006)	0.070 (0.006)	0.070 (0.006)	

NOTE: For each scenario and treatment arm, we display the average of the estimates' response rate across 10,000 simulations. We also report the standard deviation of the response rate estimates across simulations in parenthesis.

Table S2. Comparisons of the fixed randomized and response-adaptive design for MOTIVATE trials

Scenario	Average sample size and standard deviation per arm					
	Control	Arm 1	Arm 2	Arm 3	Arm 4	Total
Sample size for the original MOTIVATE design with N=6,300 enrollments						
Fixed randomization 10:2:2:3:3	3150 (39)	630 (24)	630 (24)	945 (28)	945 (28)	6,300
Sample size for the Bayesian response adaptive trial with one interim analyses						
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)	2,266 (388)	852 (271)	789 (255)	791 (256)	533 (209)	5231 (1,172)
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)	1,906 (538)	641 (379)	480 (191)	479 (190)	480 (191)	3,985 (1,065)
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)	2490 (337)	952 (198)	950 (198)	951 (199)	952 (198)	6,295 (800)
Power for the Bayesian adaptive design and fixed randomized design (in parentheses)						
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)		0.99 (0.90)	0.79 (0.80)	0.79 (0.80)	0.04 (0.05)	
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)		0.99 (0.90)	0.04 (0.05)	0.04 (0.05)	0.04 (0.05)	
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)		0.84 (0.80)	0.84 (0.80)	0.84 (0.80)	0.84 (0.80)	
Proportion of time that a trial arm is ended early for futility						
Scenario 1: (0.65, 0.75, 0.70, 0.70, 0.65)		0.00	0.04	0.04	0.56	
Scenario 2: (0.65, 0.75, 0.65, 0.65, 0.65)		0.00	0.56	0.57	0.56	
Scenario 3: (0.65, 0.70, 0.70, 0.70, 0.70)		0.04	0.04	0.04	0.04	

Note: We first determined the sample average sample size of the fixed randomized trial design (assuming n=6,300) to achieve an arm-specific power of 80% for response rate of 0.7 of an experimental arm. We then considered the power and sample size of Bayesian adaptive trials with identical average sample size design of 6,300.

SUPPLEMENT REFERENCES

1. Yokum D, Lauffenburger JC, Ghazinouri R, Choudhry NK. Letters designed with behavioural science increase influenza vaccination in Medicare beneficiaries. *Nature Human Behaviour.* 2018;2(10):743-749.
2. Choudhry NK, Krumme AA, Ercole PM, et al. Effect of Reminder Devices on Medication Adherence: The REMIND Randomized Clinical Trial. *JAMA internal medicine.* 2017;177(5):624-631.
3. Choudhry NK, Krumme AA, Ercole PM, et al. Rationale and design of the Randomized Evaluation to Measure Improvements in Non-adherence from Low-Cost Devices (REMIND) trial. *Contemporary clinical trials.* 2015;43:53-59.
4. Trippa L, Lee EQ, Wen PY, et al. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol.* 2012;30(26):3258-3263.
5. Rothman KJ. No adjustments are needed for multiple comparisons. *Epidemiology.* 1990;1(1):43-46.
6. Parker RA, Weir CJ. Non-adjustment for multiple testing in multi-arm trials of distinct treatments: Rationale and justification. *Clin Trials.* 2020;17(5):562-566.
7. Sanfelix-Gimeno G, Franklin JM, Shrank WH, et al. Did HEDIS get it right? Evaluating the quality of a quality measure: adherence to beta-blockers and cardiovascular outcomes after myocardial infarction. *Medical care.* 2014;52(7):669-676.
8. Lauffenburger JC, Franklin JM, Krumme AA, et al. Predicting Adherence to Chronic Disease Medications in Patients with Long-term Initial Medication Fills Using Indicators of Clinical Events and Health Behaviors. *J Manag Care Spec Pharm.* 2018;24(5):469-477.