bmjmedicine

Designing and conducting adaptive trials to evaluate interventions in health services and implementation research: practical considerations

Julie C Lauffenburger [1,2], Niteesh K Choudhry [1,2], Massimiliano Russo [2,3], Robert J Glynn,[2,4] Steffen Ventz,[5] Lorenzo Trippa[5]

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmed-2022-000158).

[1]Center for Healthcare Delivery Sciences, Brigham and Women's Hospital, Boston, MA, USA
[2]Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, MA, USA
[3]Department of Data Science, Dana-Farber Cancer Institute, Boston, MA, USA
[4]Division of Preventive Medicine, Brigham and Women's Hospital, Boston, MA, USA
[5]Dana-Farber Cancer Institute Department of Biostatistics and Computational Biology, Boston, MA, USA
Correspondence to: Dr Niteesh K Choudhry, Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital, Boston, MA, USA; nkchoudhry@bwh.harvard.edu

Cite this as: *BMJMED* 2022;1:e000158. doi:10.1136/bmjmed-2022-000158

Received: 7 February 2022
Accepted: 14 June 2022

Randomised controlled clinical trials are widely considered the preferred method for evaluating the efficacy or effectiveness of interventions in healthcare. Adaptive trials incorporate changes as the study proceeds, such as modifying allocation probabilities or eliminating treatment arms that are likely to be ineffective. These designs have been widely used in drug discovery studies but can also be useful in health services and implementation research and have been minimally used. In this article, we use an ongoing adaptive trial and two completed parallel group studies as motivating examples to highlight the potential advantages, disadvantages, and important considerations when using adaptive trial designs in health services and implementation research. We also investigate the impact on power and the study duration if the two completed parallel group trials had instead been conducted using adaptive principles. Compared with traditional trial designs, adaptive designs can often allow the evaluation of more interventions, adjust participant allocation probabilities (eg, to achieve covariate balance), and identify participants who are likely to agree to enrol. These features could reduce resources needed to conduct a trial. However, adaptive trials have potential disadvantages and practical aspects that need to be considered, most notably: outcomes that can be rapidly measured and extracted (eg, long term outcomes that take considerable time to measure from data sources can be challenging), minimal missing data, and time trends. In conclusion, adaptive designs are a promising approach to help identify how best to implement evidence based interventions into real world practice in health services and implementation research.

KEY MESSAGES

⇒ Adaptive trials are increasingly emerging as options to increase the efficiency and scale of interventions that could be tested in clinical medicine; while they could be used in translation of healthcare delivery interventions, they have had limited use in this context
⇒ In particular, adaptive trials could be well suited for health services research and implementation research approaches that drop inferior arms, adjust allocation probabilities, adjust sample size, or study multicomponent interventions
⇒ Simulation studies indicate that adaptive designs for two parallel group trials can have advantages over conventional, fixed non-adaptive designs (including decreasing required sample sizes), the length of the trial, and the precision of effect estimates, depending on the outcome measurement window
⇒ Adaptive trials will be more difficult to use in settings where outcomes are not rapidly retrievable or measurable from the data sources, where substantial data are missing, and when there are significant time trends

Introduction

In conventional fixed randomised controlled trials, participants are randomised to treatment groups and followed until outcomes are evaluated, generally using intention-to-treat principles. While these designs are widely considered the preferred method for evaluating the efficacy and effectiveness of healthcare interventions,[1] their limitations have been well described.[1] Most notable among these limitations is their relative inefficiency.[2–4] In many traditional randomised controlled trials, in health services and implementation research, interventions to be tested are set at the beginning of the study, and regardless of what happens during the course of the study, neither treatment assignment or allocation probabilities are modified.[2]

By contrast, in adaptive randomised trials, outcomes are observed and analysed at prespecified interim time points and modifications to study design can be made based on these observations, including modifying randomisation strategies or dropping inferior treatment arms (figure 1).[5] Adaptive multiarm designs might require fewer patients than traditional randomised controlled trials[6] and could allow for the testing of multiple interventions with more efficiency, but they also have important caveats, most notably increasing trial and methodological complexity.[7] The most common types of adaptive trial designs (in order) include: phase 2/3 studies that combine phase 2 and 3 trials, adaptive group sequential trials (which use interim stopping rules), biomarker adaptive trials (which adapt according to biomarkers), adaptive dose finding studies (which adjust allocation probabilities), pick-the-winner or drop-the-loser design (which drops inferior arms), and sample size re-estimation (which adjusts sample size based on interim data).[6 8] Other types of adaptive designs and variations of these existing ones have also been used.[9]

While adaptive trials have been widely used in early phase clinical studies, particularly in oncology,[10–15] they also appear well suited for research domains further along the translational research spectrum. Implementation science and health services research studies often seek to identify the most effective

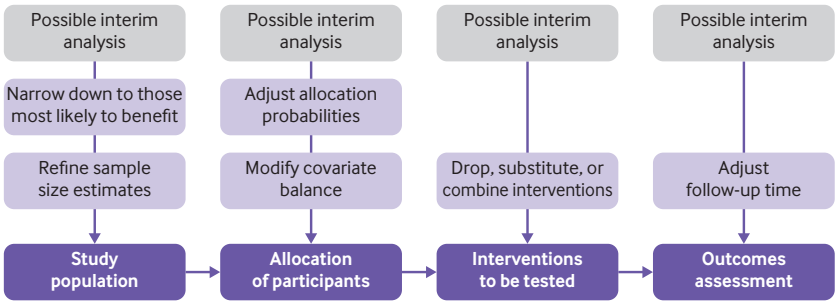

Figure 1 | Overview of potential adaptive design options through interim analyses

intervention, policy, or tool among a wide variety of possible strategies. Yet their use in these contexts remains extremely limited. Trials in this field typically evaluate healthcare delivery interventions for their real world effectiveness on health outcomes. As such, of the most common adaptive trial designs, those that adjust allocation probabilities, drop inferior arms, or adjust sample size would be particularly helpful in this type of research.[16 17]

In this article, we describe the potential advantages, disadvantages, and important considerations when applying these types of adaptive trials to implementation and health services research. We specifically consider: which interventions can be tested with adaptive designs; how eligibility criteria, enrolment procedures, and allocation probabilities can be modified; what outcomes can be evaluated and conducting interim analyses; and what the implications are on trial sample sizes and length of follow-up. Each consideration is outlined in table 1 and described in further detail throughout.

Motivating examples

To illustrate advantages, disadvantages, and considerations of adaptive designs in implementation and health services research, we describe three motivating case examples: an ongoing, pick-the-winner adaptive randomised pragmatic trial of healthcare delivery interventions and two completed multiarm pragmatic trials. We also evaluate the two completed trials' operating characteristics had they been conducted using an adaptive design that prospectively adjusts allocation probabilities during the trial. We focused on this type of design as illustration, because it is the most common adaptive design of those we believe could be useful in this field.[8]

Case example 1: NUDGE-EHR (adaptive, randomised pragmatic trial)

The NUDGE-EHR (Novel Uses of adaptive Designs to Guide provider Engagement in Electronic Health Records) study is a two stage, 16 arm, adaptive randomised trial with a pick-the-winner design that seeks to identify the most effective electronic health record tools for reducing prescribing of high risk drug treatments (NCT04284553).[18] In stage 1, 201 primary care providers were randomised to usual care (81 providers) or in equal proportions to one of 15 electronic health record tools designed using behavioural principles (eight providers/arm; online supplemental figure S1). After an eight month follow-up, arms were ranked by their impact on prescribing (ie, discontinuation or tapering drug treatments of interest), using electronic health record data. The five best performing interventions were then selected. In stage 2, usual care providers in stage 1 were randomised in equal proportions to one of the selected arms or to continue to receive usual care; and stage 1 providers who were in one of the unselected arms were re-randomised in equal proportions to a selected arm or usual care.[18]

Case example 2: MOTIVATE (fixed randomisation pragmatic trial)

The MOTIVATE (Mail Outreach To Increase Vaccination Acceptance Through Engagement) trial was a five arm, parallel group, pragmatic randomised trial testing whether the incorporation of behavioural science into mailed communication increased rates of influenza vaccination (NCT02243774).[19] Here, 228 000 Medicare beneficiaries were randomly assigned to control (no contact) or to one of four active arms in which participants received letters that all included information about vaccination but which varied the signatory and prompts. Letter 1 was from the National Vaccine Programme Office; letter 2 was from the US Surgeon General; letter 3 was from the US Surgeon General

Table 1 | Overview of practical considerations for adaptive trials in health services or implementation research

Trial characteristics	Considerations in adaptive trials
Interventions appropriate for testing	► Individual intervention components should be easily described ► Interventions should be able to be allocated to study participants over different time periods ► Multicomponent interventions can be studied if components are tested across multiple arms
Eligibility criteria, enrolment procedures, and allocation probabilities	► Narrow down recruitment of participants to those most likely to benefit in subsequent trial stages ► Address covariate balance by adjusting sample composition to improve balance
Choice of outcomes and interim analyses	► Primary outcome for adaptation is ideally measured quickly ► Outcome must be rapidly retrievable from underlying data sources ► Multiple interim analyses typically done, which can be modified based on follow-up
Required sample size and length of follow-up	► Follow-up time may be until enough outcome data points are measured, rather than an a priori window ► Using long-term outcomes for adaptation could greatly extend the length of the trial ► Substantial missing data or loss-to-follow-up issues may lengthen necessary follow-up time

and contained an implementation intention prompt (ie, asking patients to complete a plan for receiving the vaccine); and letter 4 was from the US Surgeon General and contained an active choice enhanced prompt for implementation (ie, asking patients to choose between completing a plan for the vaccine). The primary outcome was a binary outcome of influenza vaccination receipt in the four month follow-up, measured using insurance claims.

Case example 3: REMIND (fixed randomisation pragmatic trial)

The REMIND (Randomised Evaluation to Measure Improvements in Nonadherence from low cost Devices) trial was a four arm, parallel group, pragmatic randomised controlled trial that tested whether simple devices improved adherence to drug treatments (NCT02015806). This trial allocated 22 163 participants using drug treatments for cardiovascular or another non-depression condition in a 1:2:2:2 allocation ratio to control (no contact) or to receive one of three devices designed to help adherence. These devices included a strip with buttons to be toggled after taking each day's dose, a digital timer cap, and a standard daily pillbox. The primary outcome was optimal adherence over 12 months after randomisation (binary outcome), measured using insurer claims.[20 21]

Comparing fixed and response adaptive randomized designs

To illustrate implications for sample size, study duration, and power that could be obtained using adaptive trial strategies, we compare the operating characteristics of the MOTIVATE and REMIND trials using the original design or outcome adaptive bayesian design. We considered three outcome scenarios with varying effectiveness across arms (table 2). These scenarios were selected because they were thought to have at least one case favouring each type of randomised controlled trial (ie, adaptive or fixed/conventional). Scenario 1 assumed different treatment effects across intervention arms. Conversely, scenario 2 would be favourable to an adaptive randomised controlled trial because one intervention arm was designed to be clearly more effective another arms, and scenario 3 would favour fixed randomisation because no intervention arm was superior to another intervention arm. We also varied the outcome measurement windows and number of interim analyses.

We chose these examples to demonstrate the impact of effectiveness, lengths of time needed to measure outcomes (ie, short term *v* long term) and number of interim analyses on sample sizes and trial duration. We otherwise used original trial assumptions for all scenarios. For MOTIVATE, we assumed that the 228 000 participants were allocated to five treatment arms in a fixed 10:2:2:3:3 ratio for the non-adaptive design, and that controls had a 65% vaccination rate. We then considered relative effect sizes of 5-10% compared with control. For REMIND, we assumed an allocation ratio of 1:2:2:2 for the non-adaptive design, 22 163 participants, a 2% rate of adherence in the control arm,[20] and relative effect sizes of 5-8% versus control.

For further detail, see online supplemental section S1. Simulation findings, including estimated average sample sizes in each arm and average duration of the entire trials, are in table 2 and described throughout the following sections. For these simulations, we assumed that interim analyses would take 30 days; in practice, the time to conduct them can vary depending on the trial and trial oversight, although to our knowledge, this has not been precisely calculated and published.

Interventions appropriate for testing

Compared with traditional randomised controlled trials that include all interventions at the trial outset, adaptive trials can add, drop, or combine interventions on the basis of the results of interim analyses (ie, pick-the-winner or drop-the-loser designs). Interventions to be tested using adaptive designs are ideally assigned sequentially or discretely. In implementation and health services research, this means that interventions such as educational and counselling sessions, technologies, health insurance benefit structures, or checklists with discrete components are likely more suitable for adaptive trials. Accordingly, different devices (eg, REMIND) or different letters (eg, MOTIVATE) could be tested sequentially using adaptation. However, some interventions would be less ideal, notably those that cannot be applied sequentially. For example, interventions where all participants in one group are exposed simultaneously, such as an organisational change, would eliminate the ability to adapt.

Moreover, multicomponent interventions, such as those combining financial incentives with other modes of patient engagement, can also be tested in adaptive trials if intervention subcomponents are suitable for randomisation. Accordingly, investigators could reduce the number of interventions that continue to be tested over time. Adaptive trials could also allow researchers to evaluate individual parts of multicomponent interventions using regression modelling with each component being a factor in the analysis.[18] Statistical models of the effects and interactions of single interventions allow the selection and prioritisation of promising multicomponent interventions. While researchers might not have as much power to evaluate individual components as studying the multicomponent intervention, when it is of particular interest to understand the impact of individual components, this approach might prove useful.

Table 2 | Changes in sample size for MOTIVATE and REMIND trials if designed as adaptive trials under varying assumptions

Scenario (outcome probabilities for each arm in the trial in order, with control listed first)	Mean (SD) sample size per arm						Mean trial duration (months)
	Control	Arm 1	Arm 2	Arm 3	Arm 4	Total	
MOTIVATE trial							
Original trial design	114 002 (238)	22 798 (143)	22 798 (143)	34 205 (169)	34 198 (173)	228 000	4
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)	59 170 (1319)	25 878 (458)	25 876 (460)	25 875 (458)	7611 (87)	144 411 (87)	13
Scenario 2 (0.65, 0.75, 0.65, 0.65, 0.65)	39 457 (12 088)	13 661 (11 969)	7628 (150)	7627 (140)	7628 (135)	76 001 (0)	8.6
Scenario 3 (0.65, 0.70, 0.70, 0.70, 0.70)	74 464 (6424)	38 382 (1614)	38 385 (1615)	38 383 (1614)	38 385 (1617)	228 000 (0)	26
Shorter outcome measurement window and one 30 day interim analysis							
Scenario 1 (0.65, 0.75, 0.70, 0.70, 0.65)	71 451 (7407)	21 782 (2472)	21 786 (2471)	21 782 (2473)	11 401 (95)	148 201 (95)	2.5
Scenario 2 (0.65, 0.75, 0.65, 0.65, 0.65)	59 187 (18 148)	20 612 (18 195)	11 401 (95)	11 399 (95)	11 402 (95)	114 001 (0)	3.2
Scenario 3 (0.65, 0.70, 0.70, 0.70, 0.70)	87 689 (11 893)	35 074 (2978)	35 080 (2978)	35 078 (2976)	35 079 (2979)	228 000 (0)	5
REMIND trial							
Original trial design	3165 (53)	6332 (66)	6332 (67)	6333 (67)	NA	22 163	12
Same outcome measurement window as original trial and two 30 day interim analyses							
Scenario 1 (0.02, 0.10, 0.07, 0.02)	5280 (673)	3967 (844)	2275 (785)	202 (768)	NA	12 724 (278)	35.6
Scenario 2 (0.02, 0.10, 0.02, 0.02)	5258 (715)	5002 (985)	1202 (768)	1203 (768)	NA	12 665 (0)	35.4
Scenario 3 (0.02, 0.07, 0.07, 0.07)	5346 (793)	2554 (1034)	2534 (1029)	2543 (1027)	NA	12 976 (1507)	36.3
Shorter outcome measurement window and one 30 day interim analysis							
Scenario 1 (0.02, 0.10, 0.07, 0.02)	6050 (1690)	1940 (655)	1932 (655)	1753 (598)	NA	11 675 (884)	6.8
Scenario 2 (0.02, 0.10, 0.02, 0.02)	5821 (1791)	1754 (598)	1753 (598)	1753 (598)	NA	11 081 (0)	6.5
Scenario 3 (0.02, 0.07, 0.07, 0.07)	6120 (1848)	1946 (816)	1949 (822)	1950 (823)	NA	11 964 (2316)	7

MOTIVATE=Mail Outreach To Increase Vaccination Acceptance Through Engagement; REMIND=Randomised Evaluation to Measure Improvements in Nonadherence from low cost Devices; SD=standard deviation; NA=not applicable. For each scenario, mean sample sizes per arm and corresponding standard deviations across 10 000 simulations are reported, comparing the original MOTIVATE and REMIND trials' study designs with a bayesian adaptive study design (see online supplemental section S1).

For instance, NUDGE-EHR used a pick-the-winner design to eliminate ineffective interventions after the first interim analysis and used a design to evaluate multicomponent interventions (eg, a follow-up message in addition to decision support).[18] Similarly, had the MOTIVATE trial used an adaptive design, relevant letter features could have been compared, and based on interim analyses, resources could be concentrated on letters combining the most promising features. Additionally, for REMIND, ineffective devices could have been eliminated, as in scenario 1 where arm 4 was eliminated after the first interim analysis, or stage 1 (table 2).

Of note, interventions are not widely combined after interim analyses in adaptive trials presently, which is at least in part because multicomponent interventions are rarely evaluated in drug development. Conversely, many health services and implementation research trials evaluate multicomponent interventions; for instance, in studies aimed at drug treatment adherence, most effective interventions are multicomponent.[22 23]

Eligibility criteria, enrolment procedures, and allocation probabilities

To reduce required resources for a trial, adaptive trials can alter eligibility criteria to increase the likelihood of enrolment, adjust recruitment methodology, and modify allocation probabilities. This concept is similar to evaluating the degree of reach, or representativeness of participants willing to participate as an implementation outcome.[24 25] So, in addition to altering which interventions are being tested, adaptive trials can change the ratio of participants allocated to each arm in subsequent trial stages (eg, from 1:1 to 2:1). An example is shown in MOTIVATE simulation scenario 1 where more individuals received effective treatments than they would have received in the original scenario (table 2; eg, 25 878 *v* 22 798 participants in arm 1 and 59 170 *v* 114 002 participants in the control arm).

Accordingly, in health services and implementation research, the most common data sources used are electronic health records, administrative claims, and self-report.[23 26–28] For researchers wishing to adapt enrolment or eligibility criteria, electronic health records or claims data would be most helpful because they are routinely collected and would allow for adaptation without needing patient interaction. This would also reduce the amount of missing data when evaluating differences between those individuals who do and do not participate.[25] For example, if REMIND had been adaptive, baseline characteristics of patients who agreed to use the devices (such as the number of drug treatments in their regimen) could have been used to change subsequent enrolment.

In adaptive trials, allocation probabilities can be easily altered in trials that enrol on a rolling basis. For trials that enrol participants all at the same time, such as REMIND or MOTIVATE, the recruitment strategy would need to be modified.[1 29] In other words, adaptive trials can be designed to identify characteristics of those responsive to interventions and adjust accordingly (ie, a population enrichment design). For example, in REMIND, men responded better than women to the pillbottle strip; accordingly, if REMIND had been adaptive, subsequent stages after interim analyses could have preferentially allocated male patients to the pillbottle strip.

Allocation to treatment arms could also be adjusted to improve balance on baseline participant factors, such as sociodemographic characteristics, which might be particularly relevant for studies in health services or implementation research. Making such adjustments is understandably easier in trials with rolling recruitment but is also possible with simultaneous enrolment. For example, in MOTIVATE, the percentage of patients who receive the influenza vaccination in the previous season differed slightly between arms (ie, arm 5 had the lowest rate compared with the other arms), and while this imbalance was controlled for in modelling, a higher proportion of previously vaccinated patients could have been assigned to arm 5 to improve balance after interim analyses. Care must be taken for accounting for potential changes in the risk of the outcome for participants enrolled over time; if charges are large, these time trends could be a disadvantage of adaptive trials.

Choice of outcomes and interim analyses

In adaptive trials, the primary outcomes used for adaptation in interim analyses need to be rapidly retrievable for evaluation.[10] For health services and implementation research, the need for rapid retrieval means that an information system is needed to capture the outcome for evaluation in as near to real time as possible. The time needed for the interim analyses itself largely depends on the complexity of the data being collected and on data vetting before interim decisions. Of the main data sources used in this field, the lag time for accessing administrative claims data can sometimes preclude their use in interim analyses, particularly for medical claims, which take months to fully adjudicate; pharmacy claims are complete within days and might be more easily used.[27] To access information on emergency room visits or hospital admission outside of medical claims, researchers might consider using admission-discharge-transfer feeds within electronic health records, which are also available quickly. Of course, researchers or practices receiving these data through agreements might experience further delays but this process would be less problematic for trials conducted and evaluated within insurance systems.[27 30]

By contrast, electronic health records data are recorded in real time and can be retrieved as soon

as data are recorded.[31] They might therefore be useful for rapidly observing outcomes in adaptive trials. In NUDGE-EHR, adaptation occurred based on prescribing data from electronic health records. Other outcomes useful for adaption could include biometric data, such as blood pressure, weight, specific lab test results (eg, glycated haemoglobin A1c), or receipt of laboratory tests or preventive screenings such as colonoscopies. Similarly, patient reported information from connected devices might also provide data appropriate as outcomes for adaptive trials in this field.[26] Examples include accelerometer data from smartphones about physical activity, home blood pressure or glucose monitors, especially if delivered via Bluetooth connections that require minimal patient manipulation and can be quickly acquired. Their increasing real time data availability could provide avenues for future adaptive trials.

However, missing data can pose a substantial issue for adaptive trials, which is typically more problematic for patient reported outcomes or those that require patient follow-up in person (eg, cholesterol).[32 33] This is a potential disadvantage of using adaptive trials in health services research, given that many trials are designed to be pragmatic and therefore could have higher rates of missing data than drug development studies.[34] Similarly, high rates of dropout or loss to follow-up, while challenging for any trial, might actually pose a greater challenge for adaptive trials because missing data can therefore produce biased parameter estimates and participant allocation. Thus, choosing outcomes (eg, prescribing or ordering or presence or absence of diagnoses) that are less subject to having missing data could be particularly important for trials in this space. Of note, although the outcomes chosen for the NUDGE-EHR, MOTIVATE and REMIND simulations were binary, outcomes used for adaptation might take other forms, such as being continuous, depending on data completeness.

It is possible to incorporate other slower data sources where outcomes cannot be measured rapidly in an adaptive trial. In NUDGE-EHR, while the outcome for adaptation is prescribing that is assessed using electronic health record data, long term outcomes including medication filling and all cause hospital admissions can be assessed at the end using administrative claims data. Put into context, the length of the outcome (described in the next section) might have a greater influence on the ability to conduct an adaptive trial than being able to measure outcomes rapidly, but both are important considerations.

Finally, a common question for adaptive trials is how many interim analyses should be conducted. In practice, interim analyses most commonly occur one to three times throughout adaptive trials, because of cost and duration.[8] Increasing the number of interim analyses typically does not translate into dramatic gains in efficiency and accuracy of final findings. In fact, more frequent evaluations of data to stop interventions for futility can decrease power of the final analysis.[35] Also, when most participants are already enrolled, it becomes difficult to improve efficiency based on interim decisions, unless investigators allow for variations of the individual intervention and longitudinal modelling of outcomes over time.[36] Other researchers have also provided calculations that help reduce the delay in interim analyses where, for example, recruitment is continued during interim analyses.[37]

Required sample size and length of follow-up

The overall sample size can be adjusted during interim analyses to ensure desired power when effect estimates are different than originally contemplated (ie, a sample size re-estimation design).[10] Although difficult in practice, sample size re-estimation could in principle also be used to refine the intraclass correlation for cluster randomised trials.[38] However, given that cluster trials are commonly used in this field and that intraclass correlations contribute substantially to underpowered trials, this remains an area of interest.[38]

In MOTIVATE and REMIND, reductions in necessary sample size depended on the effectiveness of the interventions but suggested that sample sizes could have been smaller for each adaptive approach compared with traditional approaches when some interventions were more effective than others (table 2). For example, scenario 2 in MOTIVATE had the smallest relative total sample size (30% of the original trial) because one arm was much more effective than others. Scenario 3 in MOTIVATE was no different than the original in expected sample size because each active arm had the same effectiveness, and thus on average no differential allocation was possible during interim analyses.

The precision of effect estimates for each scenario for an adaptive trial that adjusts allocation probabilities is shown in online supplemental table S1. For MOTIVATE, this scenario suggests modest changes in precision despite smaller overall sample sizes for scenarios 1-3. For REMIND, because of the smaller overall sample size, we observed increased standard errors in scenario 1-3. This loss of precision in the estimates, however, did not translate in a loss of power in the overall trial decisions. When we repeated scenarios 1-3 with a shorter follow-up time, the length of follow-up would reduce precision, although still require smaller sample sizes than the original trial. Of note, the differences between MOTIVATE and REMIND in relative changes in sample size are largely due to baseline assumptions about the rate of outcomes in the control arms. For consistency, we chose to use the original assumptions and power calculations; operating characteristics would differ (and suggest larger necessary

sample sizes) if recalculated based on actual trial results, which showed smaller effect sizes. We also considered additional power calculations for the MOTIVATE trial with a substantially reduced sample size (ie, n=6300) to reduce the potential influence of large sample sizes on precision. We tuned the bayesian adaptive design to have a similar average sample size and power as fixed randomisation when all the interventions share the same effectiveness (scenario 3). In these simulations, when some interventions are more promising than others (scenarios 1 and 2), the considered bayesian adaptive design has higher power with a lower average sample size (online supplemental table S2).

Similarly, if NUDGE-EHR had been a 16 arm, parallel group, non-responsive adaptive trial, we estimate that >1 50 000 patients (from about 6000 physicians) would be needed to achieve equivalent power (ie, probability of detecting positive effects with significant findings) under the same assumptions.

One important limitation of adaptive trials is that the use of long term outcomes for adaptation could greatly extend the trial duration. In MOTIVATE and REMIND, the primary outcomes (influenza vaccination receipt over four months and drug treatment adherence over 12 months, respectively) were relatively long term outcomes. Thus, if the outcomes in the simulation were measured over the same window as in the original trials, an adaptive design would extend time needed for the trial (table 2).

Alternatively, if the outcome measurement windows were modified as in table 2, the trial length could be preserved while reducing sample size. For example, in REMIND, a 12 month follow-up was used, being a common interval for adherence studies.[23 39] However, the timeframe could have been compressed with outcome differences being measured over a shorter time frame (eg, three months) as shown in table 2. Furthermore, these are average durations and therefore could be shorter, especially with only one interim analysis and in situations where one arm is superior. As a result of setting the duration of the REMIND trial to match the original trial (12 months), more participants were randomised to the control arm; regardless, the overall necessary sample size was still lower. Of course, modifying the outcome measurement windows could affect the ability to observe the estimated effect size, so this may not always be appropriate.

Summary of potential strengths and limitations of adaptive trials

Adaptive trials are increasingly emerging as options to increase the efficiency and scale of interventions tested in clinical medicine. These designs could also be more widely used to potentially support more efficient evaluation and translation of healthcare delivery interventions. To our knowledge, previous work has not illuminated the extent to which adaptive trials could be specifically applied within implementation and health services research. Approaches that adjust allocation probabilities, drop inferior arms, or adjust sample size might also be well suited for health services and implementation research. The adaptive implementation of the MOTIVATE and REMIND trials suggest some advantages of adaptive trials but also illustrate potential disadvantages, including an impact on precision and potential for increasing average trial duration.

When considering other types of trials that also allow for the testing of numerous interventions, adaptive trials have several advantages. For example, factorial designs have been used in implementation research[40 41] but have their own limitations, such as that all combinations of interventions studied must be implemented and having more than two intervention factors can be complex. Sequential Multiple Assignment Randomised Implementation Trial (SMART) designs are also increasingly being used, yet fundamentally are a special case of factorial designs involving multistage randomisations to modify the intervention for participants who already received the intervention if the first stage intervention was unsuccessful.[41 42] Unlike adaptive designs, SMART designs do not adjust overall sample size or allocate new participants to study arms, unlike the broader set of possible adaptive trials.

Adaptive trials still have hugely important disadvantages in health services and implementation research, most notably the need for rapidly measurable and retrievable outcomes to not substantially increase the length of the trial. In addition, high rates of participant dropout or using outcomes for interim analyses susceptible to missing data might create more problems than in traditional randomised trials because they could lead to biased estimates of effectiveness and allocation probabilities. Similarly, considerable temporal trends can also be an limitation for adaptive designs, but some solutions do exist to resolve this and not all adaptive designs are affected equally, although designs that adjust allocation probabilities might be less suited in this scenario.[9 43] When enrolment of study participants is simultaneous or even very fast, adaptative trials will also provide substantially less usefulness. Finally, when using outcome adaptive randomisation, researchers should establish the operative characteristics of the design by conducting simulations of the trial under a set of meaningful possible cases.[44]

Conclusion

Leveraging adaptive trials for health services and implementation research could present unique opportunities to improve public health, rigor, and conduct of pragmatic trials, and more rapidly facilitate delivery of optimal healthcare. Even in health services and implementation research settings, conducting randomised trials is expensive, so

identifying ways to rigorously evaluate interventions faster will enhance the translation of evidence based interventions into real world practice.

Twitter Julie C Lauffenburger @jlauffen

Contributors JCL and NKC had overall responsibility for the design and drafted the manuscript. MR and SV contributed to the design and interpretation of the manuscript and conducted the simulation analyses and revised the manuscript for intellectual content. RJG contributed meaningfully to the design and interpretation of the manuscript and revised the manuscript for intellectual content. LT contributed to the design and interpretation of the manuscript and supervised the review and simulations. All authors approved the final manuscript. The corresponding author attests that all listed authors meet authorship criteria and that no others meeting the criteria have been omitted. JCL is the guarantor of the study.

Funding Research reported in this publication was supported by the National Institute on Aging of the National Institutes of Health (award numbers R33AG057388 and P30AG064199). JCL was also supported by a career development grant (K01HL141538) from the National Institutes of Health. LT and SV received support from the National Institutes of Health grant R01LM013352. The funders had no role in considering the study design or in the collection, analysis, interpretation of data, writing of the report, or decision to submit the article for publication.

Competing interests All authors have completed the ICMJE uniform disclosure form at www.icmje.org/disclosure-of-interest/ and declare: support from the National Institute on Aging of the National Institutes of Health for the submitted work; no financial relationships with any organisations that might have an interest in the submitted work in the previous three years; no other relationships or activities that could appear to have influenced the submitted work.

Patient and public involvement Patients and/or the public were not involved in the design, or conduct, or reporting, or dissemination plans of this research.

Provenance and peer review Not commissioned; externally peer reviewed.

Data availability statement Data are available upon request. Data will be available on reasonable request, pending appropriate agreements and institutional review board approval.

ORCID iDs
Julie C Lauffenburger http://orcid.org/0000-0002-4940-4140
Niteesh K Choudhry http://orcid.org/0000-0001-7719-2248
Massimiliano Russo http://orcid.org/0000-0003-4953-9341

REFERENCES

1 Brown CH, Ten Have TR, Jo B, *et al*. Adaptive designs for randomized trials in public health. *Annu Rev Public Health* 2009;30:1–25. doi:10.1146/annurev.publhealth.031308.100223
2 Sanson-Fisher RW, Bonevski B, Green LW, *et al*. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am J Prev Med* 2007;33:155–61. doi:10.1016/j.amepre.2007.04.007
3 Spieth PM, Kubasch AS, Penzlin AI, *et al*. Randomized controlled trials - a matter of design. *Neuropsychiatr Dis Treat* 2016;12:1341–9. doi:10.2147/NDT.S101938
4 Hartford A, Thomann M, Chen X, *et al*. Adaptive designs: results of 2016 survey on perception and use. *Ther Innov Regul Sci* 2020;54:42–54. doi:10.1007/s43441-019-00028-y
5 Wason JMS, Trippa L. A comparison of Bayesian adaptive randomization and multi-stage designs for multi-arm clinical trials. *Stat Med* 2014;33:2206–21. doi:10.1002/sim.6086
6 Hatfield I, Allison A, Flight L, *et al*. Adaptive designs undertaken in clinical research: a review of registered clinical trials. *Trials* 2016;17:150. doi:10.1186/s13063-016-1273-9
7 Pallmann P, Bedding AW, Choodari-Oskooei B, *et al*. Adaptive designs in clinical trials: why use them, and how to run and report them. *BMC Med* 2018;16:29. doi:10.1186/s12916-018-1017-7
8 Bothwell LE, Avorn J, Khan NF, *et al*. Adaptive design clinical trials: a review of the literature and ClinicalTrials.gov. *BMJ Open* 2018;8:e018320. doi:10.1136/bmjopen-2017-018320
9 Burnett T, Mozgunov P, Pallmann P, *et al*. Adding flexibility to clinical trial designs: an example-based guide to the practical use of adaptive designs. *BMC Med* 2020;18:352. doi:10.1186/s12916-020-01808-2
10 Bhatt DL, Mehta C. Adaptive designs for clinical trials. *N Engl J Med* 2016;375:65–74. doi:10.1056/NEJMra1510061
11 Bretz F, Koenig F, Brannath W, *et al*. Adaptive designs for confirmatory clinical trials. *Stat Med* 2009;28:1181–217. doi:10.1002/sim.3538
12 Trippa L, Lee EQ, Wen PY, *et al*. Bayesian adaptive randomized trial design for patients with recurrent glioblastoma. *J Clin Oncol* 2012;30:3258–63. doi:10.1200/JCO.2011.39.8420
13 Trippa L, Alexander BM. Bayesian baskets: a novel design for biomarker-based clinical trials. *J Clin Oncol* 2017;35:JCO2016682864. doi:10.1200/JCO.2016.68.2864
14 Trippa L, Rosner GL, Müller P. Bayesian enrichment strategies for randomized discontinuation trials. *Biometrics* 2012;68:203–11. doi:10.1111/j.1541-0420.2011.01623.x
15 Ventz S, Trippa L. Bayesian designs and the control of frequentist characteristics: a practical solution. *Biometrics* 2015;71:218–26. doi:10.1111/biom.12226
16 McCoy AB, Waitman LR, Lewis JB, *et al*. A framework for evaluating the appropriateness of clinical decision support alerts and responses. *J Am Med Inform Assoc* 2012;19:346–52. doi:10.1136/amiajnl-2011-000185
17 Kairalla JA, Coffey CS, Thomann MA, *et al*. Adaptive trial designs: a review of barriers and opportunities. *Trials* 2012;13:145. doi:10.1186/1745-6215-13-145
18 Lauffenburger JC, Isaac T, Trippa L, *et al*. Rationale and design of the novel uses of adaptive designs to guide provider engagement in electronic health records (NUDGE-EHR) pragmatic adaptive randomized trial: a trial protocol. *Implement Sci* 2021;16:9. doi:10.1186/s13012-020-01078-9
19 Yokum D, Lauffenburger JC, Ghazinouri R, *et al*. Letters designed with behavioural science increase influenza vaccination in Medicare beneficiaries. *Nat Hum Behav* 2018;2:743–9. doi:10.1038/s41562-018-0432-2
20 Choudhry NK, Krumme AA, Ercole PM, *et al*. Effect of reminder devices on medication adherence: the remind randomized clinical trial. *JAMA Intern Med* 2017;177:624–31. doi:10.1001/jamainternmed.2016.9627
21 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. *Contemp Clin Trials* 2015;43:53–9. doi:10.1016/j.cct.2015.05.006
22 Viswanathan M, Golin CE, Jones CD, *et al*. Interventions to improve adherence to self-administered medications for chronic diseases in the United States: a systematic review. *Ann Intern Med* 2012;157:785–95. doi:10.7326/0003-4819-157-11-201212040-00538
23 Conn VS, Ruppar TM. Medication adherence outcomes of 771 intervention trials: systematic review and meta-analysis. *Prev Med* 2017;99:269–76. doi:10.1016/j.ypmed.2017.03.008
24 Harden SM, Gaglio B, Shoup JA, *et al*. Fidelity to and comparative results across behavioral interventions evaluated through the RE-AIM framework: a systematic review. *Syst Rev* 2015;4:155. doi:10.1186/s13643-015-0141-0
25 Gaglio B, Shoup JA, Glasgow RE. The RE-AIM framework: a systematic review of use over time. *Am J Public Health* 2013;103:e38–46. doi:10.2105/AJPH.2013.301299
26 Bourke A, Dixon WG, Roddam A, *et al*. Incorporating patient generated health data into pharmacoepidemiological research. *Pharmacoepidemiol Drug Saf* 2020;29:1540–9. doi:10.1002/pds.5169

27 Choudhry NK, Shrank WH. Implementing randomized effectiveness trials in large insurance systems. *J Clin Epidemiol* 2013;66:S5–11. doi:10.1016/j.jclinepi.2013.03.022
28 Cole AP, Friedlander DF, Trinh Q-D. Secondary data sources for health services research in urologic oncology. *Urol Oncol* 2018;36:165–73. doi:10.1016/j.urolonc.2017.08.008
29 Chow S-C, Chang M. Adaptive design methods in clinical trials - a review. *Orphanet J Rare Dis* 2008;3:11. doi:10.1186/1750-1172-3-11
30 Choudhry NK. Randomized, controlled trials in health insurance systems. *N Engl J Med* 2017;377:957–64. doi:10.1056/NEJMra1510058
31 Barker PW, Heisey-Grove DM. Ehr adoption among ambulatory care teams. *Am J Manag Care* 2015;21:894–9.
32 Flory JH, Roy J, Gagne JJ, *et al*. Missing laboratory results data in electronic health databases: implications for monitoring diabetes risk. *J Comp Eff Res* 2017;6:25–32. doi:10.2217/cer-2016-0033
33 Newgard CD, Lewis RJ. Missing data: how to best account for what is not known. *JAMA* 2015;314:940–1. doi:10.1001/jama.2015.10516
34 Loudon K, Treweek S, Sullivan F, *et al*. The PRECIS-2 tool: designing trials that are fit for purpose. *BMJ* 2015;350:h2147. doi:10.1136/bmj.h2147
35 Lachin JM. A review of methods for futility stopping based on conditional power. *Stat Med* 2005;24:2747–64. doi:10.1002/sim.2151
36 Flight L, Arshad F, Barnsley R, *et al*. A review of clinical trials with an adaptive design and health economic analysis. *Value Health* 2019;22:391–8. doi:10.1016/j.jval.2018.11.008
37 Wason JMS, Brocklehurst P, Yap C. When to keep it simple - adaptive designs are not always useful. *BMC Med* 2019;17:152. doi:10.1186/s12916-019-1391-9
38 Lake S, Kammann E, Klar N, *et al*. Sample size re-estimation in cluster randomization trials. *Stat Med* 2002;21:1337–50. doi:10.1002/sim.1121
39 De Geest S, Zullig LL, Dunbar-Jacob J, *et al*. ESPACOMP medication adherence reporting guideline (emerge). *Ann Intern Med* 2018;169:30–5. doi:10.7326/M18-0543
40 Baker TB, Smith SS, Bolt DM, *et al*. Implementing clinical research using factorial designs: a primer. *Behav Ther* 2017;48:567–80. doi:10.1016/j.beth.2016.12.005
41 Brown CH, Curran G, Palinkas LA, *et al*. An overview of research and evaluation designs for dissemination and implementation. *Annu Rev Public Health* 2017;38:1–22. doi:10.1146/annurev-publhealth-031816-044215
42 Collins LM, Nahum-Shani I, Almirall D. Optimization of behavioral dynamic treatment regimens based on the sequential, multiple assignment, randomized trial (smart). *Clin Trials* 2014;11:426–34. doi:10.1177/1740774514536795
43 Villar SS, Rosenberger WF. Covariate-adjusted response-adaptive randomization for multi-arm clinical trials using a modified forward looking Gittins index rule. *Biometrics* 2018;74:49–57. doi:10.1111/biom.12738
44 Thall PF, Wathen JK. Practical Bayesian adaptive randomisation in clinical trials. *Eur J Cancer* 2007;43:859–66. doi:10.1016/j.ejca.2007.01.006

► Additional supplemental material is published online only. To view, please visit the journal online (http://dx.doi.org/10.1136/bmjmed-2022-000158).

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

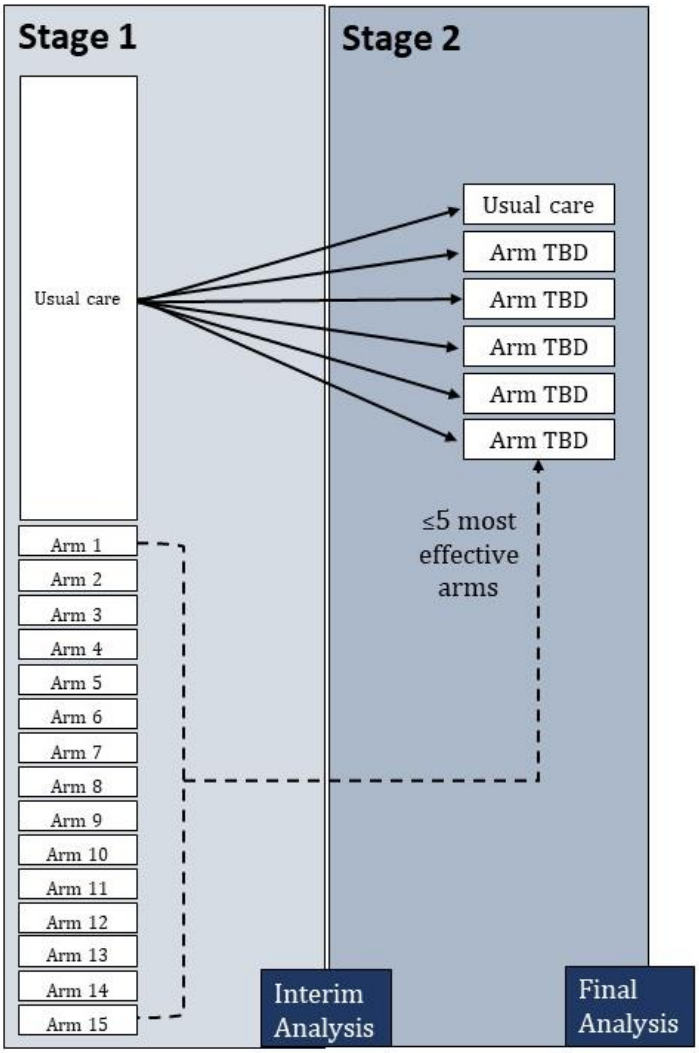

Table S1. Precision of the estimated response rates for the REMIND and MOTIVATE trials.

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.