Implementation Science

DEBATE

Open Access

Rollout trial designs in implementation research are often necessary and sometimes preferred

Gregory E. Simon[1*], Bryan R. Garner[2], Justin D. Smith[3], Peter A. Wyman[4], Theresa E. Matson[1], Lia Chin-Purcell[5], Ian Cero[4], Wouter Vermeer[6], Kimberly A. Johnson[7], Guillermo Prado[8] and C. Hendricks Brown[6]

Abstract

Background Rollout designs, which include stepped wedge designs, are defined by staggered implementation of new or alternative programs or services. Critiques of stepped wedge and other rollout designs have raised concerns regarding the confounding of true implementation or program effects with unrelated, global changes in service delivery, with some recommending they only be used when traditional parallel-group designs are not practicable. However, rollout designs may sometimes be more suitable than traditional parallel group designs for ethical, scientific, or practical reasons.

Results As investigators involved in several recent rollout trials, we define and provide rationale for and examples of stepped wedge and the larger class of rollout designs, in which all participating units receive a new program or service implementation. Staged implementation in a rollout design may be necessary when denying, rather than delaying, implementation of a known effective service is ethically unacceptable. Scientifically, stepped wedge has increased statistical power relative to an equivalent parallel group design, and some rollout designs have the capability to compare different phases of implementation and sustainment. A rollout design may be practically necessary either because of limited resources and other logistical challenges or community requirements that no site serve as a control. Examples of completed and ongoing rollout trials illustrate how these ethical, scientific, and practical considerations influenced trial designs.

Conclusions Stepped wedge and other rollout trial designs may be well suited to evaluation of implementation strategies or policy changes. In implementation trials, rollout designs may be necessary for practical reasons, may be required for ethical reasons, and may be preferred for scientific reasons. We summarize when such rollout designs have advantages and drawbacks.

*Correspondence:
Gregory E. Simon
gregory.e.simon@kp.org
Full list of author information is available at the end of the article

Contributions to the literature

- Rollout designs, which include stepped-wedge designs, are defined by staggered implementation of a new program or service across clusters or groups.
- Rollout designs may be necessary for practical reasons (especially acceptability to research partners), may be required for ethical reasons, and may be preferred for scientific reasons (including greater statistical power).
- Choice of a specific rollout design depends on the specific practical, ethical, and scientific considerations of a new service or program and the service setting where it will be implemented.

Introduction

Clinical trial designs involving random assignment of individual patients or service users into parallel groups are often poorly suited to program evaluation or implementation science research, because programs or implementation strategies are typically applied to groups, such as clinics, agencies, or geographic areas. Randomizing individuals is less efficient for evaluating programs delivered to groups, such as clinics [1]. Group randomized trials avoid this particular concern, but assigning half of the groups as controls throughout the study may be unacceptable to community partners. In contrast, rollout implementation trials [2], including stepped wedge implementation trials, assign all groups or sites to eventually receive an intervention, starting at a randomly assigned time. That staggered implementation can both increase acceptability to partners and make best use of limited implementation resources. There are concerns in using rollout trial designs in implementation research. In particular, critiques of stepped wedge designs [3] have raised concerns regarding the confounding of true implementation or program effects with unrelated, global changes in population health or service delivery. According to this and other critiques [3, 4], stepped wedge and other rollout designs may sometimes be necessary for practical reasons (such as acceptability to participants), but should be used only when more traditional parallel-group trial designs are not practicable. We argue instead that the choice of a specific design depends on a confluence of scientific considerations, ethical obligations, and practical constraints and note that stepped wedge and other rollout designs may provide ethical, scientific, and logistical advantages compared to other designs. As investigators involved in several recent rollout trials, we describe the range of rollout trials and derive principles for when they have advantages and drawbacks in implementation research. Examples from the field of implementation science illustrate these principles.

Terminology

Research designs for implementation research include the planned set of procedures to test a specific hypothesis by: (a) defining the conditions (e.g., implementation strategies) to be compared; (b) selecting units for study, most often groups or sites; (c) assigning units to conditions/time (or observe their naturally occurring assignments); and (d) assess relevant outcomes before, during, and after assignment in the conduct of the study [2, 5]. Table 1 defines specific terms useful in the description of rollout designs.

The broad category of rollout designs is defined by staggered or staged implementation of new or alternative programs or services. Groups or clusters cross over from one condition of interest to another, at one or more transition points or rollouts. With an implementation rollout design we can evaluate a novel implementation strategy for an existing or new treatment, program or service, and compare this to an existing or other novel implementation strategy. Figure 1 illustrates different types of rollout designs. As described below, categories of rollout designs vary in how groups or clusters are allocated to timing of rollout or crossover and especially in how much that allocation is under the control of the investigator or evaluator.

Stepped wedge designs, as originally defined [6], are a subset of rollout designs in which all units or clusters cross over or change to the same new condition in well-defined steps (Fig. 1a). Relevant outcomes (e.g., implementation reach [7]) are measured repeatedly in all units before and after that crossover or rollout, thus forming two complete wedges. All stepped wedge designs have this basic structure of fixed steps that form cohorts of units with the same crossover time, thus separating two complete wedges of times for each cohort where new as well as old conditions occur. Because all units contribute observations before and after their crossover time, analyses can leverage both between-cohort comparison (comparing at each step units that have and have not yet rolled out the new implementation strategy) and within-unit comparison (comparing time before and after rollout for any individual site). Rollouts at multiple well-defined steps help to disentangle true effects of crossover from general time trends during the study period [6]. As originally defined, stepped wedge designs may or may not include random assignment of units or clusters to rollouts or timing of cross-over.

Variations on the classic stepped wedge design could include (1) adding a second crossover point to a new condition, (2) diverse ways to assign when sites receive

Table 1 Glossary of terms relevant to rollout designs in implementation research

Term	Definition	Examples
Trial Design	A specification of hypotheses and how they are to be addressed rigorously in a comparison of alternatives	Implementation trial that uses a stepped-wedge design to compare implementation as usual to a new implementation strategy
Condition	A specified state defining a testable hypothesis	New implementation strategy
Unit	A generic member of a named entity or level in a design	Patient, Clinician, Clinic
Unit of observation	Level of unit where observation occurs	Program reach measured at patient level; fidelity measured at clinician level
Unit of assignment	Level of unit where implementation assignment occurs	Clinics assigned to different implementation strategies; counties assigned to different time of implementation
Assignment	Process or instantiation of assigning units to conditions	Random assignment of clinics to start implementing at different times, Selection of implementation timing by agency leadership
Cluster	A collection of units sharing one or more attributes	All clients served by the same case manager
Pre-existing Clusters	Units with an affiliation that predates the design	All clinicians working in the same clinic
Design-formed Clusters	Units grouped together by the design itself	A learning collaborative formed by having clinics work together to overcome implementation barriers
Phase	A defined section regarding the course of a specific implementation strategy	A sustainability phase when external supports are withdrawn from an implementation strategy
Stage	A section of an implementation strategy with clearly defined events for entrance and completion	Engagement stage of the Stages of Implementation Completion
Implementation Outcome	A measure of the implementation quality, quantity, speed, or duration	Fidelity to the defining or necessary aspects of a program or service
Cohort	A cluster of those units assigned to start their designated new implementation at same time	All clinics in a stepped-wedge design that start delivering a new implementation strategy at the same time
Roll-Out or Step	A point in time where an initial assignment of one or more units occurs	A stepped-wedge design with 6 distinct times where one or more units begins implementing a new strategy
Period	A segment of time where all units' implementation strategies are held fixed	The time interval between two roll-outs
Wedge	An upper or lower triangular structure, cross-classified by time periods and cohorts, all sharing the same implementation strategy and same measurement	In a design that begins measuring implementation outcomes when a unit switches to its new implementation strategy, the triangle consisting of time points where this occurs
Randomized Block Design	Units within the same cluster are assigned randomly to alternative conditions	A wait-listed design where units are first paired to form a block, then assigned randomly to which begins a new implementation strategy first
Randomized Roll-Out Design	Units are randomly assigned their start times for a new condition	A randomized stepped-wedge implementation trial
Head-to-Head Implementation Trial	Two conditions are compared by random assignment of conditions within a cohort	A single-wedge roll-out design where each cohort divides units into two starting implementation strategies and begins measuring outcomes then
Stepped-Wedge Design	All cohorts cross from one condition to another, with measurement of outcomes before and after crossover in each unit	A 2-wedge stepped wedge design where patient-level reach is measured the same way across all time periods and cohorts

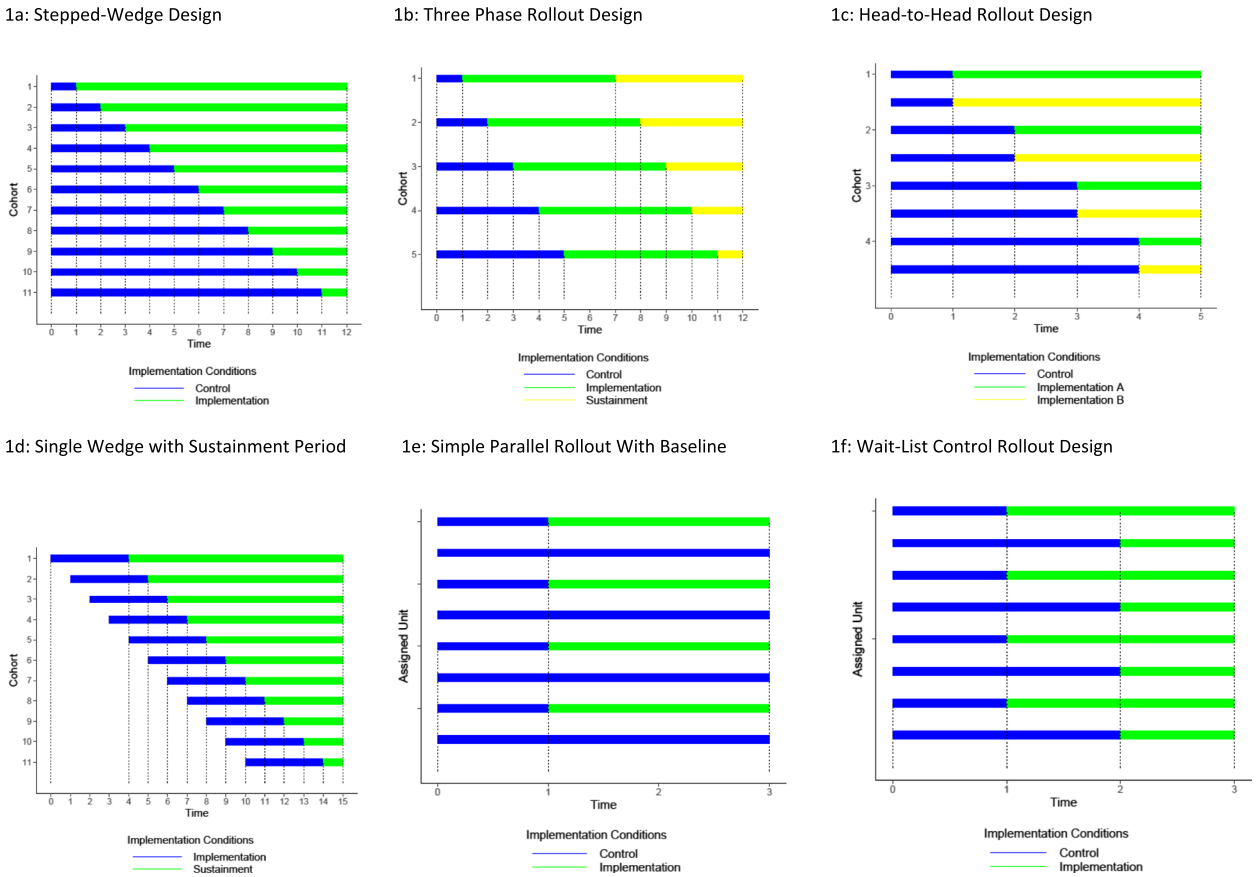

Fig. 1 Variations on rollout designs in implementation research

different conditions, (3) initiating outcome measurement only after implementation or crossover, and (4) flexible timing of when steps occur. For an example of a trial with a second crossover to a new condition (1), three-phase rollout designs (Fig. 1b) begin with implementation as usual, transition to a new active implementation strategy, and typically add a sustainment period when implementation support is discontinued or reduced after the previous period ends. In this variation, analyses can examine learning effects (gradual gains in implementation outcomes during the implementation support phase) and gradual deterioration effects after support is withdrawn. For an example of different ways to assign conditions (2), consider a head-to-head rollout trial (Fig. 1c), where sites within the same time cohort are assigned to one of two distinct implementation strategies, thus allowing a direct comparison. For an example of a design with delayed measurement of implementation outcomes (3), a single wedge rollout design (Fig. 1d) measures implementation outcomes only following the rollout of a new program or implementation strategy. This design is appropriate when evaluating a new program or service that was completely unavailable or not allowed prior to rollout so that measurement of implementation prior to rollout would be uninformative. For an example of flexible timing of steps (4), consider sites that transfer to the new condition only when they have achieved a pre-specified criterion that is considered necessary before implementing the new condition. Any of these more general rollout designs are variations on a stepped wedge design that could include random assignment of units to cohorts or, in the case of head-to-head designs, to alternative programs or intervention strategies.

The broader family of rollout designs includes a wide variety of strategies for staging of implementation. The simplest rollout design, a parallel design with baseline assessment, includes a single crossover or rollout time, with a portion of the sites crossing over to a new program or implementation strategy and the remainder continuing as usual (Fig. 1e). A variation on that simplest design includes a subsequent crossover or rollout for the remainder (Fig. 1f), which is a classic wait-listed design. In a pairwise enrollment rollout design (Fig. 1f) pairs of sites are formed in sequence, with one site allocated immediately to a new implementation and one site allocated to retain its implementation under usual conditions. Each

of these rollout design variations allows the possibility of full random assignment to when the crossover occurs. In all rollout designs, clusters may be randomly allocated to timing of rollout or crossover. In general, random assignment supports stronger causal inference regarding impact of new programs or implementation strategies. As discussed below, however, random assignment may sometimes be unacceptable to study participants or may reduce generalizability by limiting participation to those willing or able to accept random assignment.

Rollout implementation designs vs. traditional perspective of individual and cluster-randomized parallel randomized trials

Rollout designs involve specific departures from traditional individually randomized, parallel-group trials.

First, rollout designs allocate implementation of programs or services at a group or cluster level (but individual patient or service user level may be used for evaluating effectiveness in hybrid studies) [8]. Analyses must therefore account for those group and cohort effects. Implementation, policy, or program evaluation research typically focuses on effects at the level of clinician, facility, organization, or governmental entity; they can also include examination of equity via variation in reach by patient or service user characteristics [9].

Second, all rollout designs involve allocated groups crossing over to the implementation condition. Traditional parallel-group randomized trials create comparison conditions or counterfactuals using between-individual or between-group comparisons, with each individual or cluster assigned to one condition or the other. In contrast, all rollout designs utilize not only on between site comparisons but also within-site before-after comparisons.

Third, investigators evaluating implementation strategies or policy changes may have less control over the specific timing of implementation or crossover for each cluster or unit. Some rollout designs may involve random allocation, but (as discussed below) random assignment may not be ethical, acceptable, or able to be held fixed for all sites throughout the trial. Because rollout designs are conducted in real-world settings, investigators may need to accommodate changing policy or implementation decisions during the study period. More sophisticated rollout designs are well suited to flexibly evaluating implementation effectiveness across changing policy or organizational environments.

Ethical obligations

Implementation and care improvement research almost always address how programs or services that have already demonstrated effectiveness can be best implemented or delivered. Systematically denying those known effective services, as in a traditional randomized trial, may be ethically unacceptable. Staggered implementation, which only delays availability for some patients or service users, is often a more acceptable alternative. Delaying a potentially beneficial program may be more ethically acceptable than denying it completely. This concern is most relevant when comparing the current state to an implementation strategy involving added resources and less relevant when comparing alternative new implementation strategies with likely benefit. Investigators' or program developers' enthusiasm regarding a new program does not imply that withholding implementation support is unethical. Ethical equipoise depends on evidence rather than belief. But even when a new program or service is proven effective, rigorous comparison to current state may be necessary to accurately estimate incremental cost or broader benefits.

Randomly assigning groups or clusters to different times of intervention can be an equitable approach to allocating limited resources. If immediate implementation is not practicable or affordable, randomly assigning the timing or order of allocation may be an ethically acceptable approach. There may be benefits for those units who go first, as they can anticipate that the evidence-based intervention can benefit those they serve, and those units that start later they may also benefit from knowledge gained from earlier units with an improved implementation delivery.

Scientific considerations

Individually randomized clinical trial designs are often held up as the gold standard for causal inference, particularly when they achieve high rates of intervention fidelity and participant retention [10]. Parallel group randomized trials, which are more appropriate than individually randomized designs for implementation, can also provide sound causal inferences, although they require stronger assumptions to overcome an RCT's conventional "stable unit treatment value assumption" (the assumption that a unit's response depends on the treatment it receives and not the treatment of other units) [11]. Rollout trial designs for implementation can support valid inference regarding program effects, particularly when four specific types of biases are controlled: enrollment bias, assignment bias, condition biases, and external factor bias. First, enrollment bias refers to the sites that are included in a design. If they do not represent the planned target population of sites, say those serving low-income populations, and the design excludes clinics serving low-income patients, such as Federally Qualified Health Centers (FQHCs) in the US, then no amount of analysis will allow this to have external validation. Second, assignment bias

refers to how these sites are assigned to steps in rollout trial. As these designs compare implementation outcomes before and after crossover to a different implementation condition, we need to exclude other time-related bias explanations of these differences by conditions. Assignment bias is minimized first when timing of sites' transition to a new implementation condition involves balancing on important site level covariates across time (i.e., equivalent cohorts), and random assignment to timing. Third, condition bias refers to variations in how sites are examined over time. If measures taken after crossover are collected with a different instrument than before crossover, any observed differences could be due to this measurement bias, for example. Finally, external factor bias refers to changes that affect all sites differently over time. Because the new implementation condition occurs after the crossover period in rollout trials, the proportion of sites in the new implementation condition is always increasing in time, therefore external factors, such as the introduction of global policy changes, may differentially affect the two conditions. There are both design and analytic strategies for rollout designs to account for some if not all these biases. Rollout designs are also subject to some of the biases common to all trial designs, including measurement bias and selective recording of results.

Prevention or care improvement programs often operate at a group or site level. In individually randomized trials, the effects of changes in clinician knowledge or behavior or in organizational culture during the course of a trial are often considered "nuisance variables", unsuitable for study and possibly obscuring the specific effects of some new drug, device, or procedure. In contrast, in implementation research, those changes in clinician or organizational behavior are essential elements of any new program, policy, or implementation strategy. They are signal rather than noise or contamination. Consequently, allocating interventions and their respective implementations and analyzing effects at the level of clinic, health system, community agency, or governmental entity is essential for accurate assessment. For this reason, it is good practice to examine whether there are changes in implementation outcomes as a function of how long the site has been receiving the condition.

In comparison to some common parallel group randomized trial designs, rollout designs can often increase statistical power or precision when all sites are measured both before and after rollout. While detailed power comparisons are available elsewhere [12], there are some general comparisons that can be made. First, let's make a fair comparison of power for a standard stepped wedge trial design (with equal durations for each step) to a two-condition parallel group randomized trial that has the same average length of follow-up as the stepped wedge trial. With N different sites in each design, the stepped wedge design's first site has one time unit in the old condition and N in the new one; the last site has N time units in the old condition and 1 in the new one while the intermediary assigned sites are more balanced. For a comparable parallel group randomized trial, half of the sites are observed in the old condition for $N+1$ times and the other half $N+1$ times in the new condition. These two designs both have $N^*(N+1) /2$ observations on either condition. Setting the new condition to increase implementation outcomes by the same amount in both designs, and setting the heterogeneity between sites, as an Intraclass coefficient (ICC) at baseline, to be the same for the two designs as well, statistical power computations show [13] that the stepped wedge design has greater statistical power than that for the parallel design whenever there is even a tiny amount of heterogeneity in the sites (i.e., an ICC greater than 0.01). This is because every site in a stepped wedge design contributes to statistical power with a within-site comparison of the two conditions, while the parallel group design only allows between site comparisons of the two conditions. In addition, the parallel design's degrees of freedom are essentially half that for the stepped wedge design [12]. (See Fig. 2).

An additional scientific value of rollout designs is that they can help explain whether an intervention shows or does not show effects on a clinical outcome in a trial. Rollout designs with implementation staggered across clusters or agencies are well-suited to addressing questions regarding practical or real-world consequences of policy changes or changes in resources. Implementation or program evaluation research often involves higher level complex systems and causal pathways compared to the biomedical mechanisms investigated in efficacy trials. Allowing that complexity to unfold across different organizations is essential to accurately assess real-world effects and maximize external validity.

Rollout designs are well-suited to evaluation of time-varying changes in program effects during and after program implementation. When implementation is staggered across groups or clusters of clinics, agencies, or geographic units, each unit can be observed during pre-implementation, active implementation, and in sustainment phases.

Implementation outcomes, as outlined by Proctor and colleagues [14] (acceptability, appropriateness, adoption, feasibility, implementation cost, penetration or reach, and sustainment) and embedded in RE-AIM [7], may be collected across phases in rollout designs to understand what works when. Implementation outcomes may be measured across sites before and during implementation rollout to get a developmental perspective using the Stages of Implementation Completion (SIC) [15].

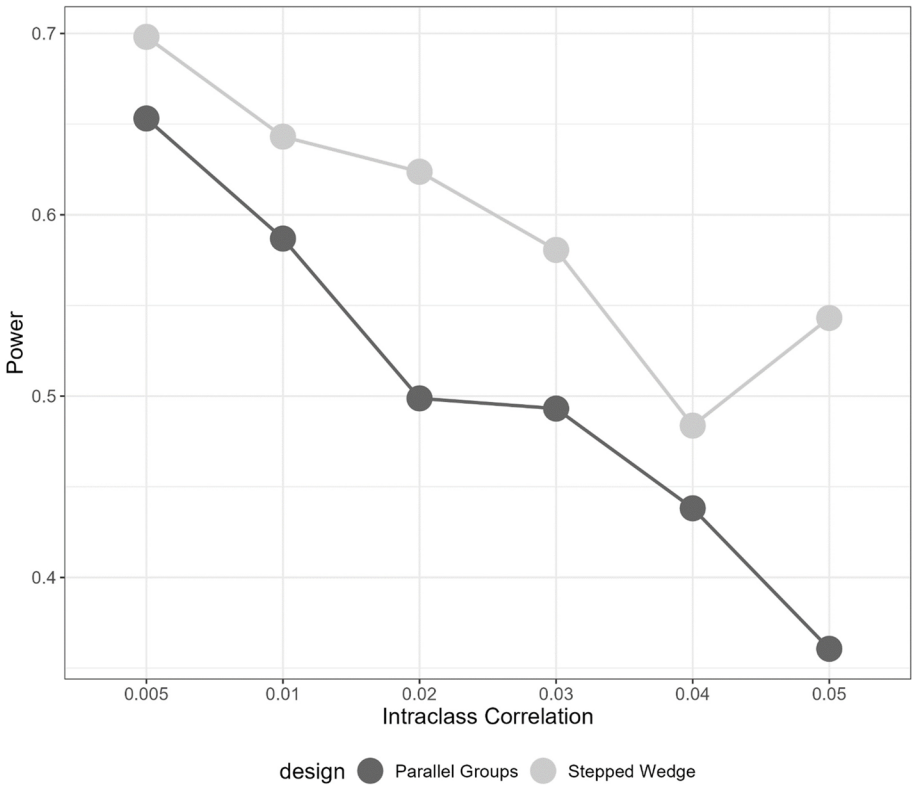

Fig. 2 Simulated power of parallel vs. stepped wedge designs at varying levels of shared variance between sites. All simulations followed the model outlined by Hussey & Hughes [24]

The SIC can distinguish when and how implementation changes occur. Reach, often defined in each of these approaches as the proportion of eligibles who receive an intended intervention, is highly relevant to the goals of implementation, typically measured at the cluster-level, and can be continuously monitored and tracked throughout a rollout design using routinely collected administrative data such as from insurance billing or electronic health records. Likewise, acceptability, appropriateness, and feasibility may be dynamic and change as a function of time and exposure to the intervention. They are well-suited to the time-varying design of rollout trials and may be measured using repeated surveys. Moreover, when an intervention is newly implemented (i.e., no baseline measurement of reach), acceptability, appropriateness, and feasibility data may be collected prior to implementation and compared to post-implementation data. Other implementation outcomes, including adoption, fidelity, and sustainment, are only measurable during specific phases of implementation, but their success or failure at one phase, such as preparation, is a good predictor of implementation success [16]. Thus, time from adoption to full implementation may be meaningful in rollout trials and can be tracked and measured using the SIC [15].

We note that some implementation strategies, such as those using learning collaboratives or networks, deliberately bring together multiple sites and therefore they can no longer be considered as independent sites for analysis. Rollout designs, as well as parallel group trials, that include random assignment to modest sized learning collaboratives can account for such non-independence [8, 9].

Practical considerations

Pragmatic trial innovations, such as waiver or alteration of individual informed consent and reliance on automatically collected data to assess outcomes, can facilitate use of traditional parallel-group trial designs in community settings. Nevertheless, some practical considerations may favor rollout over parallel-group designs.In describing those practical considerations, we first consider ways that rollout reduces resource constraints. We then consider ways that practical considerations for using rollouts can ameliorate ethical as well as scientific concerns that have already been described.

Resource constraints and logistical considerations involving implementation are often sufficiently demanding that they cannot be delivered with fidelity to all sites

at once within the typical research budget. In such cases, staggered implementation from a rollout design is often a good option [17]. For example, when initial implementation of a new program or service requires significant training, technical assistance, or external facilitation, limited availability of staff or other necessary resources may not permit simultaneous implementation across all participating units – or even simultaneous implementation at half of all units as would be done in a parallel-group randomized trial. Rollout designs with stepped wedge or other staggered implementation can be a practical approach to making optimal use of limited implementation resources.

Regarding ethical and scientific concerns for conducting implementation research, we have noted that rollout designs can provide an acceptable alternative to sites that would otherwise serve as controls. Even when investigators are in ethical equipoise (i.e. evidence does not clearly indicate superiority of an implementation strategy under study, there may be a strong desire among communities, organizations, or institutions to implement in all sites at once, especially when addressing individual outcomes as serious as overdose, suicide, and HIV transmission. Thus, acceptability of study assignment to potential participants may be a scientific consideration as well as a practical constraint. Participating clinicians, agency staff, or leaders assigned to a no-intervention condition might attempt to adopt or implement some components of the program or service under study, introducing attenuation in estimation of program effects. Alternatively, those disappointed by assignment to a control or no-intervention condition might be less engaged in contributing outcome data or report poorer outcomes due to a disappointment effect. Through a transparent process of random assignment of the sites to when they would begin implementing, key community leaders may be more willing to adhere to the randomization schedule, being assured that when it is their turn all sufficient resources would be available.

Examples

Specific studies listed below illustrate some variations of rollout trial design. For each example, choice of a specific design was informed by a combination of the three elements of ethical obligations, scientific considerations, and practical considerations.

eHealth familias unidas trial [18]

Intervention A virtually delivered version of Familias Unidas Mental Health, a family-based intervention for Hispanic youth previously shown to prevent/reduce drug use, depressive symptoms, suicide and sexual risk behaviors.

Setting Eighteen pediatric primary care clinics.

Design This three phase stepped wedge hybrid effectiveness-implementation randomized rollout trial randomly assigned clinics to implement the program at one of five rollouts or steps separated by three months. The three phases are pre-implementation (families enrolled in a control condition with no family program), Familias Unidas Mental Health implementation (a one-year period for each clinic where families are enrolled in this virtual intervention), and a sustainment period following active implementation.

Outcomes Effectiveness outcomes (mental health symptoms, substance use, family functioning) were measured and compared across all three phases. Implementation outcomes (reach, adoption, and fidelity) compared the active implementation to the sustainment phase.

Ethical considerations A traditional parallel design, assigning some clinics to provide no effective intervention, was considered not ethically acceptable.

Scientific considerations Control youth recruited in the pre-implementation phase are re-recruited in the next phase where they do receive Familias Unidas and thus support analyses of developmental changes.

Practical considerations Implementation of Familias Unidas carried out by a limited number of expert external facilitators, thus rollout spread over 15 months was practically necessary.

Northwestern University Improving the Management of symPtoms during and following Cancer Treatment (NU IMPACT) trial [19, 20]

Intervention Electronic health record-integrated symptom monitoring and management program (cPRO).

Setting Thirty two adult outpatient cancer care clinics in an academic health system.

Outcome(s) Primary implementation outcome was reach, the proportion of patients engaging with cPRO each month. Secondary implementation outcomes included reach of cPRO (proportion engaging among those eligible), reach of referral (proportion among those

eligible) and clinician adoption (proportion of clinicians using cPRO tools to respond to patient reports).

Design This Type 2 hybrid effectiveness-implementation study followed a stepped wedge design with the 32 participating clinics each randomly assigned to implement at one of seven steps or rollouts. An embedded patient-level randomized controlled trial, following a staircase design, was used to test the effectiveness of cPRO compared to usual care. Patients consented to the RCT were excluded from the implementation analysis of the stepped wedge as their participation could bias implementation outcomes.

Ethical considerations Cancer symptom monitoring is a required practice of health systems accredited by the Commission on Cancer (CoC) of the American College of Surgeons. Thus, it was not ethically acceptable to assign some clinics to not implement cPRO or some similar program.

Scientific considerations As the study was interested in testing the impact of implementation strategies on a clinic-level implementation outcome (proportion of patient adoption within clinics), as opposed to their impact on a patient-level outcome, a cluster randomized stepped wedge design was optimal for power given the number of clinic month proportions to be included in the analysis ($n=432$).

Practical considerations Due to the large number of discrete strategies comprising the multicomponent package, and the finite availability of health system staff from Operations and Quality departments for this initiative, a roll-out was more feasible than a parallel design with half of clinics implementing simultaneously.

Northwestern Collaborative Behavioral Health Project (CBHP) [21]

Intervention University of Washington's AIMS Center Collaborative Care Model (CoCM) for depression and anxiety, including systematic follow-up by a care manager, stepped-care treatment including antidepressant medication and psychotherapy, and systematic psychiatric consultation.

Setting Eleven academically affiliated adult primary care clinics in the Northwestern Medicine Central Region.

Outcome(s) Implementation outcomes (reach, adoption) measured immediately before implementation, during a 12-month period of active implementation, and during a subsequent sustainment phase. Stages of Implementation Completion (SIC) was used to evaluate pace and quantity of implementation activities across phases for each clinic.

Design Hybrid type 2 effectiveness-implementation trial following a randomized rollout design (that resembles a single wedge design), with 2 of 11 participating clinics designated as high priority for implementation were randomly assigned to either of two initial rollout steps with the remaining 9 clinics assigned to one of 9 subsequent steps using matched-pair randomization.

Ethical considerations Given evidence for effectiveness of CoCM, it was not acceptable to assign clinics to NOT implement. The roll-out design and associated implementation strategies also targeted processes that would increase equity were selected in collaboration with health system leaders.

Scientific considerations Given the small sample size of clinics, a parallel cluster-randomized design would yield little implementation outcome data and would be difficult to balance the two arms given heterogeneity in clinic size and other variables. A roll-out design with matched-pair randomization to sequence provided far more data on CoCM implementation and helped to address potential challenges with balance in the allocation.

Practical considerations Resources to support implementation were limited, and this concern was magnified during the COVID-19 pandemic when this study took place. The rollout design spread the need for implementation resources over a longer period.

A head-to-head trial of two implementation strategies for delivering multidimensional treatment foster care trial [22]

Intervention Two distinct approaches to implement Multidimensional Treatment Foster Care, an evidence-based program to support foster parents as an alternative to group care. The standard implementation supported single county implementation by external facilitators while a Community Development Team (CDT) learning collaborative delivered implementation support for the same intervention to a group of 6–8 counties at a time.

Setting Fifty-one counties in California and Ohio, requiring collaboration between child welfare, juvenile justice, and mental health service systems.

Outcomes Completion, speed, and quality of implementation measured before implementation, during active implementation, and during subsequent sustainment are obtained from the SIC.

Design Head-to-head randomized rollout design with counties randomly assigned to one of four steps or rollouts and then counties in each step randomly assigned to CDT-facilitated implementation or implementation as usual.

Ethical considerations California required all counties to deliver evidence-based alternatives to group foster care, including Multidimensional Treatment Foster Care, which could only be trained by the research developers. A rollout design was necessary to give every county this training.

Scientific considerations Counties were balanced and randomized twice to timing of implementation and whether they received the CDC or standard implementation. The head-to-head design allowed balance between these conditions across the entire trial, an essential component since a major recession occurred during the trial and affected both conditions. Analyses needed to account for the fact that CDT learning collaborative deliberately created dependence between counties in the same group.

Practical considerations Three separate steps or rollouts were necessary in California given limited resources to support implementation across all sites concurrently. Due to the slow-down of county services during the recession, Ohio counties were added to the end of the original design and randomized similarly to CDT and standard implementation.

Wingman Connect (WC) US Air Force (USAF) expansion trial [23]

Intervention Interactive, group training for early-career personnel to prevent suicide risk, depression, and occupational problems.

Setting US Air Force (USAF) technical training school and 8 bases launching implementation, followed by force-wide scale-out of the Wingman Connect (WC) Program delivered by USAF personnel.

Outcomes Implementation outcomes of implementer fidelity and engagement, examined as a function of alignment of the WC program with base leadership activities, climate and embeddedness into base communications and support activities. Effectiveness outcomes were base-level suicide attempt rates and suicide risk scores of enrolled Airmen.

Design Hybrid implementation-effectiveness (type 1) trial, with a 2-stage randomized design conducted concurrently. Stage 1 involved randomized job training classes and enrolled participants (WC vs. control) and stage 2 was a stepped wedge randomization of the order in which eight operational bases begin to implement WC for all incoming early career Airmen including those enrolled in stage 1.

Ethical considerations The 8 operational bases in the stepped wedge included sites with elevated suicide rates, and USAF leadership would not have approved randomly assigning some to not implement any prevention program.

Scientific considerations This design used all 8 bases for studying WC implementation, whereas a parallel group design would have only compared 4 implementing vs. 4 control sites. It also included complementary data from other USAF bases that were exposed to much less implementation of WC; this helped adjust for potential external changes in context regarding suicide and USAF administrative changes.

Practical considerations It was logistically impracticable to prepare all 8 bases to implement simultaneously and thus the roll-out design was practical and efficient. The 8 bases were selected because they received a large portion of enrolled participants from the technical training school, thus integrating these two concurrent trials.

Cautions regarding rollout designs

Because rollout designs all involve within-cluster comparisons of time before and after rollout, they may be more liable to confounding by external events unrelated to the program or strategy under study. External events, such as the COVID-19 pandemic, may disrupt implementation activities or service delivery. Within-cluster comparisons across time may be distorted by those disruptions or may not be generalizable to times outside of those disruptions. External events, such as the transition from the ICD-9-CM diagnosis coding system to ICD-10-CM and organizational changes in electronic medical records may disrupt measurement of implementation outcomes, making comparisons across time difficult to interpret. Head-to-head rollout trials are subject to less bias than other rollout trials. While investigators or evaluators cannot anticipate

specific disruptive events, such as the COVID-19 pandemic, they must often adapt to unanticipated events during a rollout trial. That adaptation should consider how any specific external event may affect the health condition of interest, the delivery of specific services, the use of specific implementation strategies, and the tools for measurement of implementation outcomes.

Rollout designs that include several steps or rollouts across time often require longer time for enrollment and observation than do parallel-group designs in which all units cross over at once. This may delay availability of study results and increase overall costs of completing a trial. Trials including several rollouts or steps spread over two years or more may be less liable to disruption or confounding by external changes. But, if those trials include a significant period of measurement prior to implementation as well as measurement during a post- sustainment period, the trial period may extend over four years or more.

Conclusions

Stepped wedge and other rollout trial designs may be a viable and rigorous alternative to parallel-group designs, and especially in contrast to individually randomized parallel-group designs, for evaluation of implementation strategies or policy changes. Rollout designs may be necessary for practical reasons, such as acceptability to community partners, or may be preferred for ethical reasons, such as concerns regarding systematic denial of an effective or empirically supported practice. In addition, stepped wedge or other rollout designs may be preferred for scientific reasons, including: focus on real-world impact at the group or cluster level, gain in statistical power by combining both between cluster and within-cluster comparisons, and ability to distinguish different phases of implementation or policy change (pre-implementation, active implementation, and sustainment).

While rollout designs may be more liable to confounding or bias due to external changes in population health or service delivery, investigators and evaluators can address that potential weakness through specific design choices. Staggered rollout over several steps (rather than one or two) will increase the likelihood that implementation or program effects can be accurately distinguished from unrelated external events. Random assignment of clusters to different rollout times will reduce the likelihood of biased selection of earlier or later implementation. Consistent measurement of implementation outcomes, beginning before implementation in all clusters, will help to distinguish true program effects from artifacts of changes in measurement.

Authors' contributions

GS collaborating in envisioning the paper, drafted the manuscript, and contributed substantially to revisions. BG collaborated in envisioning the paper and contributed substantially to revisions. JS collaborated in envisioning the paper and contributed substantially to revisions, especially regarding the NU IMPACT and collaborative care trials. PW contributed substantially to revisions, especially regarding the Wingman Connect trial. TM collaborated in envisioning the paper and contributed substantially to revisions. LC collaborated in envisioning the paper and contributed substantially to revisions. IC contributed substantially to revisions, especially regarding statistical power for rollout trials. WV collaborated in envisioning the paper and contributed substantially to revisions. KJ contributed substantially to revisions. GP contributed substantially to revisions, especially regarding the eHealth Familias Unidas trial. CHB collaborated in envisioning the paper and contributed substantially to revisions, especially regarding the eHealth Familias Unidas. All authors have approved the final submitted version.

Funding

This work was supported by the National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH) under Award Number P50DA054072 (Center for Dissemination & Implementation At Stanford [C-DIAS]; PI: McGovern) and U2CDA057717 (Research Adoption Support Center [RASC]; MPIs: McGovern, Becker, Brown, Becker). The content is solely the responsibility of the authors and does not represent the official position of NIDA/NIH. Additional support by grant UM1CA233035 from the National Cancer Institute, grants R01MH124718 and R01MH076158 from the National Institute of Mental Health, and award W81XWH-14–1-0322 from the US Army Medical Research Acquisition Activity.

Data availability

Not applicable.

Declarations

Ethics approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Competing interests

JD Smith is an associate editor of *Implementation Science*. C Hendricks Brown serves on the editorial board of *Implementation Science*. The authors declare that they have no other competing interests.

Author details

[1]Kaiser Permanente Washington Health Research Institute, 1730 Minor Ave. #1360, Seattle, WA 98101, USA. [2]The Ohio State University College of Medicine, Columbus, OH, USA. [3]Spencer Fox Eccles School of Medicine, University of Utah, Salt Lake City, UT, USA. [4]School of Medicine and Dentistry, University of Rochester, Rochester, NY, USA. [5]Stanford University School of Medicine, Palo Alto, CA, USA. [6]Feinberg School of Medicine Northwestern University, Chicago, IL, USA. [7]Department of Mental Health Law and Policy, University of South Florida, Tampa, FL, USA. [8]University of Miami, Miami, FL, USA.

Received: 6 December 2024 Accepted: 16 February 2025
Published online: 24 February 2025

References

1. Brown CH, Liao J. Principles for designing randomized preventive trials in mental health: an emerging developmental epidemiology paradigm. Am J Community Psychol. 1999;27(5):673–710.
2. Brown CH, Curran G, Palinkas LA, Aarons GA, Wells KB, Jones L, Collins LM, Duan N, Mittman BS, Wallace A, Tabak RG, Ducharme L, Chambers DA, Neta G, Wiley T, Landsverk J, Cheung K, Cruden G. An overview of research and evaluation designs for dissemination and implementation. Annu Rev Public Health. 2017;38:1–22 PMCID: PMC5384265.

3. Hemming K, Taljaard M. Reflection on modern methods: when is a stepped-wedge cluster randomized trial a good study design choice? Int J Epidemiol. 2020;49(3):1043–52 PMCID: PMC7394949.
4. Hughes JP, Heagerty PJ, Xia F, Ren Y. Robust inference for the stepped wedge design. Biometrics. 2020;76(1):119–30 PMCID: PMC7978491.
5. Hwang S, Birken SA, Melvin CL, Rohweder CL, Smith JD. Designs and methods for implementation research: advancing the mission of the CTSA program. J Clin Transl Sci. 2020;4(3):159–67 PMCID: PMC7348037.
6. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. BMC Med Res Methodol. 2006;6:54 PMCID: PMC1636652.
7. Glasgow RE, Vogt TM, Boles SM. Evaluating the public health impact of health promotion interventions: the RE-AIM framework. Am J Public Health. 1999;89(9):1322–7 PMCID: PMC1508772.
8. Curran GM, Landes SJ, McBain SA, Pyne JM, Smith JD, Fernandez ME, Chambers DA, Mittman BS. Reflections on 10 years of effectiveness-implementation hybrid studies. Front Health Serv. 2022;2:1053496 PMCID: PMC10012680.
9. McNulty M, Smith JD, Villamar J, Burnett-Zeigler I, Vermeer W, Benbow N, Gallo C, Wilensky U, Hjorth A, Mustanski B, Schneider J, Brown CH. Implementation research methodologies for achieving scientific equity and health equity. Ethn Dis. 2019;29(Suppl 1):83–92 PMCID: PMC6428169.
10. Butcher NJ, Monsour A, Mew EJ, Chan AW, Moher D, Mayo-Wilson E, Terwee CB, Chee ATA, Baba A, Gavin F, Grimshaw JM, Kelly LE, Saeed L, Thabane L, Askie L, Smith M, Farid-Kapadia M, Williamson PR, Szatmari P, Tugwell P, Golub RM, Monga S, Vohra S, Marlin S, Ungar WJ, Offringa M. Guidelines for reporting outcomes in trial reports: the CONSORT-outcomes 2022 extension. JAMA. 2022;328(22):2252–64.
11. Schwartz S, Gatto NM, Campbell UB. Extending the sufficient component cause model to describe the Stable Unit Treatment Value Assumption (SUTVA). Epidemiol Perspect Innov. 2012;9:3 PMCID: PMC3351730.
12. Hughes JP, Lee WY, Troxel AB, Heagerty PJ. Sample size calculations for stepped wedge designs with treatment effects that may change with the duration of time under intervention. Prev Sci. 2024;25(Suppl 3):348–55 PMCID: PMC10950842.
13. Brown CH, Hedeker D, Gibbons RD, Duan N, Almirall D, Gallo C, Burnett-Zeigler I, Prado G, Young SD, Valido A, Wyman PA. Accounting for context in randomized trials after assignment. Prev Sci. 2022;23(8):1321–32 PMCID: PMC9461380.
14. Proctor E, Silmere H, Raghavan R, Hovmand P, Aarons G, Bunger A, Griffey R, Hensley M. Outcomes for implementation research: conceptual distinctions, measurement challenges, and research agenda. Adm Policy Ment Health. 2011;38(2):65–76 PMCID: PMC3068522.
15. Saldana L, Chamberlain P, Wang W, Hendricks Brown C. Predicting program start-up using the stages of implementation measure. Adm Policy Ment Health. 2012;39(6):419–25 PMCID: PMC3212640.
16. Alley ZM, Chapman JE, Schaper H, Saldana L. The relative value of pre-implementation stages for successful implementation of evidence-informed programs. Implement Sci. 2023;18(1):30 PMCID: PMC10362770.
17. Brown CH, Wyman PA, Guo J, Pena J. Dynamic wait-listed designs for randomized trials: new designs for prevention of youth suicide. Clin Trials. 2006;3(3):259–71.
18. Estrada Y, Lozano A, Boga D, Tapia MI, Perrino T, Velazquez MR, Forster L, Torres N, Morales CV, Gwynn L, Beardslee WR, Brown CH, Prado G. eHealth Familias Unidas Mental Health: protocol for an effectiveness-implementation hybrid type 1 trial to scale a mental health preventive intervention for Hispanic youth in primary care settings. PLoS One. 2023;18(4):e0283987. PMCID: PMC10112791.
19. Cella D, Garcia SF, Cahue S, Smith JD, Yanez B, Scholtens D, Lancki N, Bass M, Kircher S, Flores AM, Jensen RE, Smith AW, Penedo FJ. Implementation and evaluation of an expanded electronic health record-integrated bilingual electronic symptom management program across a multi-site Comprehensive Cancer Center: the NU IMPACT protocol. Contemp Clin Trials. 2023;128:107171 PMCID: PMC10164083.
20. Scholtens DM, Lancki N, Hemming K, Cella D, Smith JD. Statistical analysis plan for the NU IMPACT stepped-wedge cluster randomized trial. Contemp Clin Trials. 2024;143:107603 PMCID: PMC11283938.
21. Smith JD, Fu E, Rado J, Rosenthal LJ, Carroll AJ, Atlas JA, Carlo AD, Burnett-Zeigler I, Jordan N, Brown CH, Csernansky J. Collaborative care for depression management in primary care: a randomized roll-out trial using a type 2 hybrid effectiveness-implementation design. Contemp Clin Trials Commun. 2021;23:100823 PMCID: PMC8350002.
22. Brown CH, Chamberlain P, Saldana L, Padgett C, Wang W, Cruden G. Evaluation of two implementation strategies in 51 child county public service systems in two states: results of a cluster randomized head-to-head implementation trial. Implement Sci. 2014;9:134 PMCID: PMC4201704.
23. Wyman PA, Pisani AR, Brown CH, Yates B, Morgan-DeVelder L, Schmeelk-Cone K, Gibbons RD, Caine ED, Petrova M, Neal-Walden T, Linkh DJ, Matteson A, Simonson J, Pflanz SE. Effect of the Wingman-Connect upstream suicide prevention program for air force personnel in training: a cluster randomized clinical trial. JAMA Netw Open. 2020;3(10):e2022532 PMCID: PMC7578767.
24. Hussey M, Hughes J. Design and analysis of stepped wedge cluster randomized trials. Contemp Clin Trials. 2007;28:182–91.

Publisher's Note