DOI: 10.1093/jncimonographs/lgq014

Designing Studies That Would Address the Multilayered Nature of Health Care

David M. Murray, Michael Pennell, Dale Rhoda, Erinn M. Hade, Electra D. Paskett

Correspondence to: David M. Murray, PhD, Division of Epidemiology, College of Public Health, The Ohio State University, Columbus, OH (e-mail: dmurray@cph.osu.edu).

We review design and analytic methods available for multilevel interventions in cancer research with particular attention to study design, sample size requirements, and potential to provide statistical evidence for causal inference. The most appropriate methods will depend on the stage of development of the research and whether randomization is possible. Early on, fractional factorial designs may be used to screen intervention components, particularly when randomization of individuals is possible. Quasi-experimental designs, including time-series and multiple baseline designs, can be useful once the intervention is designed because they require few sites and can provide the preliminary evidence to plan efficacy studies. In efficacy and effectiveness studies, group-randomized trials are preferred when randomization is possible and regression discontinuity designs are preferred otherwise if assignment based on a quantitative score is possible. Quasi-experimental designs may be used, especially when combined with recent developments in analytic methods to reduce bias in effect estimates.

J Natl Cancer Inst Monogr 2010;40:90–96

Cancer care encompasses a continuum of steps from screening to end-of-life care and covers a range of populations from general (eg, all women) to specific (eg, patients for whom active cancer treatment is no longer a desirable option) (1). Linking all the steps are interfaces where information and/or responsibility are exchanged. These steps and interfaces are influenced by factors at the level of the individual; family, friends, and health-care providers; the clinic or hospital; the community; and the health-care system. To address the steps and the interfaces, interventions often need to address more than one level of influence. When they do, they are examples of multilevel interventions now common in public health and medicine (2).

One of the major challenges for evaluating multilevel interventions is that patients who receive care from the same physician, clinic, or health-care system often have or develop connections through their common experiences, shared environments, or mutual interactions. These connections create a positive intraclass correlation (The intraclass correlation is the average pairwise correlation for the dependent variable among the members of the same group.) and threaten the validity of the usual analytic methods for randomized clinical trials (RCTs) (3). Application of those methods will yield a type I error rate that is inflated, often badly (4–7). When the number of groups is limited, the degrees of freedom and power for a valid test also may be limited (4,8,9). Finally, simple random assignment of a limited number of groups to each condition may not distribute all potential confounders evenly, thereby jeopardizing internal validity (8,9). Nonrandom assignment also increases the risk of confounding. Consideration must be given to all of these challenges as the study is planned and analyzed to support valid inference (8,9).

Considerable discussion has been devoted to how best to evaluate multilevel interventions (eg, 2,10). Indeed, whole conferences have been devoted to this topic (11). The discussion has not been focused on cancer care per se, but because cancer care interventions often take place on multiple levels, the discussion is quite relevant.

Some have suggested that randomized trials are either inappropriate or impractical for the evaluation of multilevel interventions (eg, 12,13). Several have offered nonrandomized alternative designs (eg, 13–18). Others have defended randomized designs and criticized the nonrandomized alternatives (eg, 19,20). Some have suggested variations on the usual randomized design (eg, 21–25).

We agree with those who have suggested that different designs are appropriate at different stages of development of the intervention (eg, 26) and under different conditions (eg, 2,27). Our purpose was to review the more promising alternatives and offer recommendations as to which may be most helpful for multilevel intervention studies across the cancer continuum.

Study Designs in the Context of Phases of Research

Flay (28) described phases in health promotion and disease prevention research relevant to this discussion, expanding on a scheme proposed earlier by Greenwald (29). We consider alternative designs appropriate to these phases.

Pilot Tests

Pilot tests evaluate the feasibility and acceptability of intervention and evaluation protocols. As such, they do not require a research design or an analysis plan to estimate intervention effects. Instead, investigators repeat the pilot testing process to refine their materials until they are ready for use in a prototype study.

Prototype Studies

Prototype studies provide preliminary testing of the intervention and evaluation materials for their effects on mediators and intermediate outcomes (30). Effects on these outcomes should be larger, and occur earlier, than effects on primary outcomes, supporting the use of smaller studies. Any of the design and analytic alternatives discussed below for efficacy and effectiveness studies could be used for prototype studies. Here, we focus on three design and analytic alternatives that appear especially promising for the evaluation of prototype studies. Several also will have application in efficacy or effectiveness studies, especially where randomization is not possible.

Fractional Factorial Designs. In the full factorial design, the factors of interest are crossed and participants are randomized to each cell in the multidimensional table defined by those factors. Analysis of variance or corresponding methods for nonnormal data are used to evaluate main effects and interactions. In a fractional factorial design, cells are selectively eliminated from the multidimensional table so as to allow evaluation of all main effects and two-way interactions but not higher-order interactions (25).

Nair et al. (25) used a fractional factorial design to screen intervention components for decision aids related to tamoxifen use among women at high risk of breast cancer. Of interest were five intervention factors, each with two conditions. A full factorial design would require 2^5 cells, and power for higher-order interactions would be quite limited without a very large sample. A fractional factorial design was used to evaluate the main effects and all two-way interactions in a design requiring only 2^4 cells and a much smaller sample. Primary outcomes were mediators and intermediate outcomes, including knowledge, perceived risk, and behavioral intentions. The investigators were able to identify components that appeared promising to investigate further in an efficacy study.

In concept, fractional factorial designs are well suited for screening intervention components during the development of any multilevel intervention. In practice, they will be most helpful when it is possible to randomize individual participants in an RCT and when the intervention effects on the mediators and intermediate outcomes are relatively large and occur soon after introduction of the intervention. The fractional factorial design also could be used with a group-randomized trial (GRT), and with interventions that have a longer latency, but that will rapidly increase the size and cost of the study and may make fractional factorial designs time- and cost-prohibitive under those circumstances. Importantly, as the latency increases, the investigator will lose the major advantage of the fractional factorial design, which is the rapid screening of intervention components.

Sample size methods for fractional factorial designs are the same as for those used for RCTs or GRTs. Fractional factorial designs are more efficient than full factorial designs in screening intervention components because they focus on main effects and two-way interactions and ignore higher-order interactions. They may offer additional savings by focusing on mediators and intermediate outcomes, which may have larger effects and shorter latency. They will have limited utility in other circumstances because designs involving more than two factors are rarely used in efficacy or effectiveness trials in cancer research.

Quasi-Experimental Designs. Quasi-experiments have as many variations as experimental designs, but the central feature is that participants are not randomly assigned (31–33). Quasi-experiments have long been recommended when randomization is not possible, whether for logistic, ethical, or other reasons. However, quasi-experiments are subject to a number of threats to internal validity that are usually well addressed by randomization, and so they are less rigorous than randomized designs (33).

Paskett et al. (34) used a quasi-experimental design to examine a cancer screening intervention. One community received an intervention to increase breast and cervical cancer screening among low-income women aged 40 years and older. A second community served as a comparison site. Cohort and serial cross-sectional data were collected in both communities at baseline and 3 years later. This quasi-experimental design assigned just one community per condition. Unfortunately, there is no valid analysis for this design without strong and untestable assumptions because variation because of community is completely confounded with variation because of study condition (35). Even in a prototype study, it would be much better to have at least two sites in each arm.

Analytic methods for quasi-experimental designs are quite similar to those used in RCTs and GRTs. Methods for sample size calculation also are quite similar. As a result, a quasi-experimental design has no inherent sample size advantage if it is powered to provide statistical evidence for effects on mediators and intermediate or primary outcomes. Moreover, the investigators also must address the additional threats to internal validity.

Recent analytic developments have improved that situation so that quasi-experimental and experimental designs can give similar results when well implemented and applied to similar problems (eg, 18,27,36,37). The best quasi-experiments will measure or have access to a rich set of covariates likely to be related both to the outcome and to the assignment. This will allow better matching of study conditions during the design (18,38) and for adjustment during the analysis (18,36).

We agree with those who have argued that a good quasi-experiment, if well analyzed, can provide strong evidence for causal inference (eg, 18,27,36,37). But we also agree with those who have argued that it is often more difficult to conduct a good quasi-experiment than to conduct a good randomized trial (19). We conclude that randomized designs are still preferred over quasi-experiments except where randomization is not possible.

Time-Series Designs. The use of time-series analysis has been investigated in the statistical literature since the early 20th century (39) and has been discussed in standard texts on quasi-experimental design for some time (eg, 32). Time-series designs (TSDs) involve repeated measurements of an outcome before and after an intervention or a policy change (ie, a change in recommended screening guidelines or a law limiting smoking in public places). Serial observations in time are usually correlated, and analytic methods used to describe or make inferences from these data need to account for that correlation. Failure to consider the correlation over time can result in underestimated standard errors and subsequent overestimation of statistical significance. The autoregressive integrated moving average model of Box and Jenkins (40) is a standard class of time-series models that can accommodate and characterize autocorrelation over time and model seasonality. Once the over time dependencies are identified, the baseline and intervention periods can be compared using simple tests to determine whether a significant change in the trend, intercept, or variability was associated with the intervention.

Michielutte et al. (41) and Goldberg et al. (42) described evaluations of cancer screening programs and physician reminder systems, respectively, to improve cancer screening rates. Michielutte et al. reported on a trend analysis of mammography screening in one public health clinic. This analysis was one part of their evaluation of a clinic- and community-based intervention program to increase cervical and breast cancer screening. As the authors describe, major limitations of their study include the lack of outcomes measured in control clinic(s) and relatively few measurements in the time series (19 data points). Conversely, Goldberg et al. presented a TSD in which two distinct geographical locations of the same physicians' practice (firms) were studied to determine whether the firm allocated by coin flip to a reminder system increased patient colorectal, breast, and cholesterol screening. Unlike Michielutte et al., Goldberg et al. included a control group but again based their evaluation on a limited number of time points.

One of the major limitations of TSDs is that for stable estimates in autoregressive integrated moving average models, 50 observations per period are recommended (32). This may be impossible or impractical in cancer care research. Another major limitation is that the single group TSD provides only a within-group comparison. The TSD can be strengthened by adding additional within-group outcomes that are not expected to change as a result of the intervention. However, even with those improvements, the investigator must rely on within-group statistical evidence and logic rather than between-group evidence for causal inference. If the investigator wants between-group statistical evidence, the number of groups required will approach that needed in a GRT. Moreover, investigators will need additional information at the planning stage beyond estimates of intraclass correlation to plan for the impact of correlation over time. The necessity for such a large number of groups and the *many* within-group observations over time is the reason we do not see between-group comparisons in time-series studies.

Multiple Baseline Designs. Multiple baseline designs have a long history in the study of individual behavior change (43). More recently, they have been advocated for the evaluation of complex multilevel interventions (13,14,17). In this design, the outcome of interest is measured repeatedly in a small number of study participants before the intervention is introduced. This allows the investigator to establish a stable baseline level for the outcome. The intervention is then introduced in one participant at a time in a random or systematic order. Once a participant moves to the intervention condition, the intervention continues for the remainder of the study. The regular measurements begun before the intervention also continue throughout the study. The investigator hopes to observe a change in the outcome in each participant following the intervention and to observe no corresponding change among the participants before the intervention. Such a pattern is taken as evidence for a causal effect; any alternative explanation would need to account for synchronicity between the intervention and the effect across participants.

Blount et al. (44) used a multiple baseline design to evaluate interventions to help pediatric oncology patients cope with painful treatment procedures. Three young children were trained in an array of distraction techniques, and their parents were trained to coach them. The intervention appeared to have the desired outcome in two of the children but not in the third.

Evaluation methods for multiple baseline designs have traditionally relied on visual comparisons of the outcome levels both within and between participants. Visual methods may suffice when the intervention has a rapid and large effect and is successful in every participating individual or group. In cases where effect sizes are modest and success is not uniform, methods of statistical hypothesis testing have been applied to both the within- and between-participant comparisons (43,45–47).

The mixed result found by Blount et al. (44) illustrates the risk of a multiple baseline design involving only a few participants. If the results are not consistent across all participants, the investigator is left to judge whether the intervention was effective absent any statistical evidence from a between-participant comparison.

The issue of sample size for multiple baseline designs has two components: the number of participants and the number of measurements for each participant. As few as two participants with appropriately synchronized interventions and outcomes might provide evidence for an intervention effect but only if the investigator is willing to rely on visual rather than statistical evidence. The sample size requirements for between-participant statistical comparisons would be similar to those for the usual RCT or GRT so that multiple baseline designs offer little advantage for those comparisons. The number of measurements required for each participant will depend on how variable the outcome is over time. More measurements will be required to establish a stable baseline if the outcome is quite variable.

Efficacy and Effectiveness Studies

Efficacy trials test whether the intervention causes the observed effect under controlled conditions. Effectiveness trials test whether the treatment will remain effective when implemented under more realistic conditions. Both efficacy and effectiveness studies require designs that can support causal inference for the primary outcome and so require a level of rigor beyond what is necessary for prototype designs.

Group-Randomized Trials. GRTs are comparative studies in which investigators randomly assign identifiable groups to conditions and observe individual members of those groups to assess the effects of an intervention (8,9). These trials and their associated analytic methods are ideally suited for efficacy and effectiveness studies of multilevel interventions because they allow for randomization at any level of influence and because they accommodate hierarchical data structures quite naturally. GRTs and RCTs are the gold standard methods in public health and medicine when randomization occurs at the group and individual levels, respectively.

Katz et al. (48) employed a GRT to evaluate a clinic-based intervention to motivate clinicians to counsel their smoking patients to quit and to offer nicotine replacement therapy to help them quit. Proactive telephone counseling also was provided to those patients. Clinics were randomized to study conditions, and patients who smoked were recruited for the study. Biochemically validated abstinence was twice as high in the intervention condition compared with the control condition. The investigators concluded that the intervention was associated with high abstinence among smokers.

Methods for sample size calculation and data analysis in GRTs are now well established (8,9,49). Analysis methods must accommodate the positive intraclass correlation expected in the data; a variety of methods can be used to do that, including mixed-model regression, permutation tests, generalized estimating equations, and two-stage analytic methods (8,9,49). Sample size requirements are greater than for RCTs because of the two penalties originally identified by Cornfield (4): extra variation and limited degrees of freedom. None of the nonrandomized approaches provides more efficient between-group statistical evidence for causal inference, though several provide non-statistical evidence in much smaller studies. GRTs are preferred when between-group statistical evidence for causal inference is required.

One of the challenges for GRTs is that they are often large and expensive studies. Their size and cost are driven by the extra variation, group-based degrees of freedom, and the complexity of the interventions they are used to evaluate. Considerable progress has been made to limit the impact of the extra variation, but little progress has been made to address the problem of limited degrees of freedom (49). Given that the complexity of the interventions is unlikely to change, the limited degrees of freedom problem stands as a good target for future methodological research.

Dynamic Wait-List or Stepped Wedge Designs. In the standard wait-list GRT, half of the groups are randomized to the intervention, whereas the other half provide control observations until the end of the study. After the final data are collected, the controls receive the intervention as compensation for their participation in the trial. The standard wait-list GRT needs only one or two measurements and often will be more efficient than other wait-list variations described below. Even so, there may be circumstances in which the standard wait-list design is not available or in which alternative designs may be more efficient.

For example, logistical or political considerations may require giving the treatment to the controls before the desired follow-up time has elapsed. Jarjoura (50) showed that if the treatment effect has a rapid onset and is stable over time, increasing the number of measurement occasions in which all participants are in the same condition (either treatment or control) will increase efficiency. That can be accomplished through multiple baseline measurements or by switching controls to treatment during the last several measurement occasions. This requires that measurements be made throughout the study or at periodic intervals and that may not be feasible.

If logistical or political considerations require that controls receive the treatment even earlier in the study, several authors have recommended giving the treatment to randomly selected controls in a staggered fashion rather than, for example, giving the treatment to all the controls halfway through the study (23,51,52). This dynamic wait-list or stepped wedge design will be more efficient when 1) it is impossible to withhold the treatment from the controls until the end of the study, 2) the outcome's sample variance decreases as the amount of time under study increases, and 3) it is possible to collect data when each new set of controls is randomly selected for treatment.

Analytic and sample size methods for these wait-list designs will be similar to those used for other GRTs. As such, these wait-list alternatives offer no sample size advantage over simpler GRTs. Indeed, the traditional wait-list design often will be the most efficient, though the Jarjoura alternative (50) may be more efficient under specific circumstances. The dynamic wait-list or stepped wedge design (23,51,52) will be less efficient but may be appropriate under certain conditions.

Additive Designs. In some circumstances, investigators are interested in the additive effects of two or more interventions or intervention components. For example, in the Child and Adolescent Trial for Cardiovascular Health (CATCH), the investigators were interested in the additive effects of a school-based intervention and a family-based intervention. They could have employed a factorial design to examine the independent and joint effects of the two interventions, but that design would have involved four sets of schools and power for the interaction effect. Power for interactions is always less than that for main effects, other factors being constant, and so factorial designs are not common in GRTs.

Instead, Child and Adolescent Trial for Cardiovascular Health employed an additive design involving three arms: control, school intervention only, school and parent intervention. This design involves only three sets of schools, rather than four, and it can be powered around any of the three pairwise comparisons.

This design also could be used to evaluate multilevel cancer interventions. For example, if the investigators were interested in examining the effects of patient, physician, and clinic interventions on cancer screening outcomes, they might consider a design with four conditions: control, patient only, patient plus physician, patient plus physician plus clinic. This arrangement would be appropriate if the other cells in the 2 × 2 × 2 = 8 cell factorial design did not make sense for logistical or political reasons, for example, or if the investigators wanted to focus power on the pairwise comparisons of these four conditions rather than on all the interactions available in the factorial design.

Regression Discontinuity Designs. Cook (53) provides an excellent history of regression discontinuity (RD) designs in psychology, education, statistics, and economics. In this design, assignment to treatment is based on a quantitative score. Those scoring on one side of a cut point receive the intervention, whereas those on the other side do not. The assignment score is then used as a covariate in the analysis of intervention effects. An advantage of this approach, compared with the traditional RCT, is that by assigning interventions based on pretest measures, the investigator can ensure that those most in need of an intervention receive it. Hence, RD designs can provide a rigorous alternative when an RCT is not possible for ethical reasons. Like fractional factorial designs, RD designs previously have been applied to individuals, but they could just as easily be applied to groups. For instance, an educational or media intervention to increase cancer screening could be implemented in those communities whose baseline screening rates fall below a prespecified threshold.

Decker (54) employed an RD design to study the effect of Medicare insurance on mammography. Using public data on the use of health-care services, breast cancer diagnosis, and survival in the United States, Decker tested for a discontinuity in these outcomes at age 65 when Medicare provides nearly universal coverage. She found a considerable drop in the percentage of uninsured patients, an increase in the percentage who had checkups and mammograms in the previous 2 years, and a modest decrease in late-stage breast cancer diagnosis at age 65. These results suggest that access to Medicare improves mammography and early detection of breast cancer. At the same time, the effect of Medicare insurance availability is completely confounded with any other phenomena that occur uniformly at age 65, so the evidence is not as strong as it would be given a randomized trial. In this case, of course, randomization is not possible, and the RD design may be the best alternative.

RD designs provide an unbiased estimate of the intervention effect when the association between the outcome and assignment variable is appropriately modeled (27,33,53). Hence, a considerable burden is placed on the analyst to choose the proper model. For this reason, some authors have explored the use of nonparametric regression in RD (eg, 55). Another practical disadvantage of RD is its decreased power relative to the RCT. Cappelleri et al. (56) demonstrate that RD designs can require more than twice the sample size of an RCT to achieve the same level of power. We are currently examining the sample size implications for interventions applied to groups, though we expect similar results. Power can be increased and inference potentially improved by combining the characteristics of RD and RCTs into the same study (56,57). For example, communities with a cancer screening rate below a lower threshold may be assigned the intervention, communities whose rate is above an upper threshold are assigned control, and those in between are randomized. Hence, this design can serve as a compromise when a randomized trial is either impossible or unethical.

Case Studies

Case 1: Follow-up of an Abnormal Mammogram

This case involves a failure to ensure follow-up of abnormal screening tests within a health center that has multiple practices. A multilevel intervention to address this problem could include three levels: 1) an organizational level that addresses the medical and administrative leadership of an organization using an academic detailing model, 2) a team level to engage members of the health-care team in adopting skills in patient-centered communication and the appropriate use of the tracking system, and 3) a patient level that includes culturally appropriate materials and instructions regarding the meaning of the test results and how patients would have their abnormal screening test evaluated. The organizational level intervention would provide the leadership with information about the screening deficits at their facility and try to elicit their support to implement a tracking system to monitor the status of individuals with abnormal screening tests.

A number of the designs could be used to evaluate the intervention at various stages of development. In a preliminary study, fractional factorial designs could be used to evaluate components at any of the three levels. Multiple components could be tested, with data collected on intermediate outcomes, such as knowledge, attitudes, intentions, and perceived barriers. Components that appeared promising would be retained for the next level of testing.

As a next step, a TSD could be employed in a single health center to provide information on whether the multilevel intervention as a whole was associated with changes in follow-up test completion. The TSD would require that records be available frequently enough to make use of the analytic methods associated with the TSD.

Alternatively, a multiple baseline design involving a few health centers could be used even if data were not available on the frequency required for the time series. Data would have to be collected periodically to establish a stable baseline, and the centers would be given the intervention sequentially and in a random order. If the pattern in the outcome was linked to the intervention, and no similar change occurred absent the intervention, the investigators would have evidence for an intervention effect.

A GRT would provide the strongest evidence but would require multiple health centers randomized to either an intervention or a control arm. The size and cost of the study could be limited by sampling practices within health centers and patients within practices and delivering the team- and patient-level interventions only to those sampled. Even if it were necessary to deliver the interventions more broadly, sampling could be used to limit the scope and cost of data collection.

An additive design could be used if there were interest in the incremental effects of the three interventions. A four-arm design might be used with a usual care control; a patient-level arm; a patient- and team-level arm; and a patient-, team-, and organization-level arm. This would be the only design of the set that would provide information on the incremental effects of the three intervention levels.

Case 2: Implementing Electronic Medical Records

The head of a large health-care organization decides that she wants to implement the electronic medical record in a way that will allow her to evaluate its impact on organizational morale, provider team functioning, and patient care. The administrative leader decides to implement it in stages among the 50 facilities within her health-care organization. She recognizes it will be a disruptive process, so she hopes to measure care and the effects of implementation in 25 clinics for a year after a 6-month implementation period.

The facilities vary in size from 5 to 15 providers serving populations from 10 000 to 30 000 patients. The populations in these clinics make an average of 40 000 visits per year to the smallest clinics and 120 000 visits per year to the largest clinics. There are five clinics with five providers, 15 with seven providers, 10 with 10 providers, and 20 with 15 providers. On average, each provider team includes a receptionist, a licensed practical nurse, and a physician. There is one nurse for every five providers. This nurse has some responsibility for quality improvement activities.

The administrator decides to emphasize staff satisfaction, teamwork, and patient outcomes she increasingly values—breast, cervical, and colorectal cancer screening rates, diabetes management, and hypertension management. Staff satisfaction and teamwork would be measured by survey. Patient cancer-related outcomes would be measured using screening rates within 2 years among patients who have been seen in the clinic at least twice in 3 years and at least once in the past 1 year. The economic levels of the populations served by these clinics differ, and some are in rural settings.

This case involves a single intervention that is expected to have effects at several levels. It could be examined using a few health centers with a multiple baseline design, but as noted above, causal inference would rely on logic rather than between-center comparisons. Alternatively, this case is a natural setting for a GRT. Clinics might be stratified based on the number of providers and the socioeconomic status of the clinic population. If possible, they could also be stratified on urban vs rural. After baseline data collection for 6 months, clinics would be randomized to intervention or control, with 25 in each arm. Data collection would continue for another 6 months during the intervention. Trends in screening and other outcomes would be compared pre- and postintervention between the intervention and control arms. This would be a very strong design and likely have very good power.

Discussion

The most appropriate design and analytic plan will depend on the stage of development of the research and whether randomization is possible. In prototype studies, which estimate effects on intermediate outcomes, fractional factorial designs may be used to screen intervention components, particularly when randomization of individuals is possible. Quasi-experimental, time-series, and multiple baseline designs can be useful in prototype studies once the intervention is designed because they require few sites and can provide the preliminary evidence for efficacy studies. In efficacy and effectiveness studies, GRTs are preferred when randomization is possible. RD designs are preferred if assignment to treatment cannot be random but can be made based on a quantitative score. Quasi-experimental designs also may be used, especially when combined with recent developments in analytic methods to reduce bias in effect estimates. Time-series and multiple baseline designs may be used for efficacy and effectiveness studies but only if the investigator is willing to rely on logic rather than between-group statistical evidence as the basis for causal inference.

References

1. Taplin SH, Rodgers AB. Toward improving the quality of cancer care: addressing the interfaces of primary and oncology-related subspecialty care. *J Natl Cancer Inst Monogr*. 2010;40:3–10.
2. Mercer SL, DeVinney BJ, Fine LJ, Green LW, Dougherty D. Study designs for effectiveness and translation research: identifying trade-offs. *Am J Prev Med*. 2007;33(2):139–154.
3. Kish L. *Survey Sampling*. New York, NY: John Wiley & Sons; 1965.
4. Cornfield J. Randomization by group: a formal analysis. *Am J Epidemiol*. 1978;108(2):100–102.
5. Zucker DM. An analysis of variance pitfall: the fixed effects analysis in a nested design. *Educ Psyc Measurmt*. 1990;50(4):731–738.
6. Murray DM, Wolfinger RD. Analysis issues in the evaluation of community trials: progress toward solutions in SAS/STAT MIXED. *J Community Psychol*. 1994;CSAP Special Issue:140–154.
7. Murray DM, Hannan PJ, Baker WL. A Monte Carlo study of alternative responses to intraclass correlation in community trials: is it ever possible to avoid Cornfield's penalties? *Eval Rev*. 1996;20(3):313–337.
8. Murray DM. *Design and Analysis of Group-Randomized Trials*. New York, NY: Oxford University Press; 1998.
9. Donner A, Klar N. *Design and Analysis of Cluster Randomization Trials in Health Research*. London, UK: Arnold; 2000.
10. Ukoumunne OC, Gulliford MC, Chinn S, Sterne JAC, Burney PGJ. Methods for evaluating area-wide and organisation-based interventions in health and health care: a systematic review. *Health Technol Assess*. 1999: 3(5).
11. OBSSR. *Workshop on Research Designs for Complex, Multi-level Health Interventions and Programs 2004*. (updated 2004; cited 2/6/09). http://obssr.od.nih.gov/news_and_events/conferences_and_workshops/FY_2004/complex_interventions.aspx.
12. Speller V, Learmonth A, Harrison D. The search for evidence of effective health promotion. *BMJ*. 1997;315(7104):361–363.
13. Sanson-Fisher RW, Bonevski B, Green LW, D'Este C. Limitations of the randomized controlled trial in evaluating population-based health interventions. *Am J Prev Med*. 2007;33(2):155–161.
14. Biglan A, Ary D, Wagenaar AC. The value of interrupted time-series experiments for community intervention research. *Prev Sci*. 2000;1(1): 31–49.
15. Gilmour S, Degenhardt L, Hall W, Day C. Using intervention times series analyses to assess the effects of imperfectly identifiable natural events: a general method and example. *BMC Med Res Methodol*. 2006; 6:16–25.
16. Glasgow R, Emmons KM. How can we increase translation of research into practice? Types of evidence needed. *Annu Rev Public Health*. 2007; 28:413–433.
17. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *Am J Prev Med*. 2007;33(2):162–168.
18. West SG, Duan N, Pequegnat W, et al. Alternatives to the randomized controlled trial. *Am J Public Health*. 2008;98(8):1359–1366.
19. Eccles M, Grimshaw J, Campbell M, Ramsay C. Research designs for studies evaluating the effectiveness of change and improvement strategies. *Qual Saf Health Care*. 2003;12(1):47–52.
20. Rosen L, Manor O, Engelhard D, Zucker DM. In defense of the randomized controlled trial for health promotion research. *Am J Public Health*. 2006;96(7):1181–1186.
21. Katz DL, Nawaz H, Jennings G, et al. Community health promotion and the randomized controlled trial: approaches to finding common ground. *J Public Health Manag Pract*. 2001;7(2):33–40.
22. Linden A, Trochim WMK, Adams JL. Evaluating program effectiveness using the regression point displacement design. *Eval Health Prof*. 2006; 29(4):407–423.
23. Brown CH, Wyman PA, Guo J, Pena J. Dynamic wait-listed designs for randomized trials: new designs for prevention of youth suicide. *Clinical Trials*. 2006;3(3):259–271.
24. 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.
25. Nair V, Strecher V, Fagerlin A, et al. Screening experiments and the use of fractional factorial designs in behavioral intervention research. *Am J Public Health*. 2008;98(8):1354–1359.
26. Campbell M, Fitzpatrick R, Haines A, et al. Framework for design and evaluation of complex interventions to improve health. *BMJ*. 2000; 321(7262):694–696.
27. Shadish WR, Cook TD. The renaissance of field experimentation in evaluating intervention. *Annu Rev Psychol*. 2009;60:607–629.

28. Flay BR. Efficacy and effectiveness trials (and other phases of research) in the development of health promotion programs. *Prev Med.* 1986; 15(5):451–474.
29. Greenwald P, Cullen JW. The scientific approach to cancer control. *CA Cancer J Clin.* 1984;34(6):328–332.
30. Stevens J, Taber DR, Murray DM, Ward DS. Advances and controversies in the design of obesity prevention trials. *Obes Res.* 2007;15(9):2163–2170.
31. Campbell DT, Stanley JC. *Experimental and Quasi-Experimental Designs for Research.* Chicago, IL: Rand McNally College Publishing Company; 1963.
32. Cook TD, Campbell DT. *Quasi-Experimentation: Design and Analysis Issues for Field Settings.* Chicago, IL: Rand McNally College Publishing Company; 1979.
33. Shadish WR, Cook TD, Campbell DT. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA: Houghton Mifflin Company; 2002.
34. Paskett ED, Tatum CM, D'Agostino RB, et al. Community-based interventions to improve breast and cervical cancer screening: results of the Forsyth County Cancer Screening (FoCaS) Project. *Cancer Epidemiol Biomarkers Prev.* 1999;8(5):453–459.
35. Varnell SP, Murray DM, Baker WL. An evaluation of analysis options for the one group per condition design: can any of the alternatives overcome the problems inherent in this design? *Eval Rev.* 2001;25(4):440–453.
36. Shadish WR, Clark MH, Steiner PM. Can nonrandomized experiments yield accurate answers? A randomized experiment comparing random to nonrandom assignment. *JASA.* 2008;103(484):1334–1346.
37. Cook TD, Shadish WR, Wong VC. Three conditions under which experiments and observational studies produce comparable causal estimates: new findings from within-study comparisons. *J Policy Anal Manage.* 2008;27(4):724–750.
38. Rubin DB. Comment: the design and analysis of gold standard randomized experiments. *JASA.* 2008;103(484):1350–1356.
39. Yule GU. On the time-correlation problem, with especial reference to the variate-difference correlation method. *J Royal Stat Soc.* 1921;84(4): 497–526.
40. Box GEP, Jenkins GM. *Time Series Analysis: Forecasting and Control.* Oakland, CA: Holden-Day, Inc.; 1976.
41. Michielutte R, Shelton B, Paskett ED, Tatum CM, Velez R. Use of an interrupted time-series design to evaluate a cancer screening program. *Health Educ Res.* 2000;15(5):615–623.
42. Goldberg HI, Neighbor WE, Cheadle AD, Ramsey SD, Diehr P, Gore E. A controlled time-series trial of clinical reminders: using computerized firm systems to make quality improvement research a routine part of mainstream practice. *Health Serv Res.* 2000;34(7):1519–1534.
43. Barlow DH, Nock MK, Hersen M. *Single Case Experimental Designs: Strategies for Studying Behavior Change.* 3rd ed. Boston, MA: Allyn and Bacon; 2009.
44. Blount RL, Powers SW, Cotter MW, Swan S, Free K. Making the system work. Training pediatric oncology patients to cope and their parents to coach them during BMA/LP procedures. *Behav Modif.* 1994;18(1):6–31.
45. Marascuilo L, Busk P. Combining statistics for multiple-baseline AB and replicated ABAB designs across subjects. *Behav Assess.* 1988;10(1):1–28.
46. Ferron JM, Bell BA, Hess MR, Rendina-Gobioff G, Hibbard ST. Making treatment effect inferences from multiple-baseline data: the utility of multilevel modeling approaches. *Behav Res Methods.* 2009;41(2):372–384.
47. Bulte I, Onghena P. Randomization tests for multiple-baseline designs: an extension of the SCRT-R package. *Behav Res Methods.* 2009;41(2): 477–485.
48. Katz DA, Muehlenbruch DR, Brown RL, Fiore MC, Baker TB. Effectiveness of implementing the agency for healthcare research and quality smoking cessation clinical practice guideline: a randomized, controlled trial. *J Natl Cancer Inst.* 2004;96(8):594–603.
49. Murray DM, Varnell SP, Blitstein JL. Design and analysis of group-randomized trials: a review of recent methodological developments. *Am J Public Health.* 2004;94(3):423–432.
50. Jarjoura D. Crossing controls to treatment in repeated-measures trials. *Control Clin Trials.* 2003;24(3):306–323.
51. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28(2):182–191.
52. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol.* 2006;6:54.
53. Cook TD. "Waiting for life to arrive": a history of the regression-discontinuity design in psychology, statistics and economics. *J Econometrics.* 2008;142(2):636–654.
54. Decker SL. Medicare and the health of women with breast cancer. *J Hum Resour.* 2005;40(4):948–968.
55. Imbens G, Lemieux T. The regression discontinuity design—theory and applications. *J Econometrics.* 2008;142(2):611–614.
56. Cappelleri JC, Darlington RB, Trochim WMK. Power analysis of cutoff-based randomized clinical trials. *Eval Rev.* 1994;18(2):141–152.
57. Trochim WMK, Cappelleri JC. Cutoff assignment strategies for enhancing randomized clinical trials. *Control Clin Trials.* 1992;13(3):190–212.

Funding

National Cancer Institute (P50CA105632).

Affiliations of authors: Division of Epidemiology, College of Public Health (DMM, EDP), Comprehensive Cancer Center (DMM, EDP), and Division of Biostatistics, College of Public Health (MP, DR, EMH), The Ohio State University, Columbus, OH; Center for Public Health Research (DR) and Center for Evaluation (DR), Battelle Memorial Institute, Columbus, OH; Center for Biostatistics, The Ohio State University, Columbus, OH (EMH).