Challenges of Prolonged Follow-up and Temporal Imbalance in Pragmatic Trials: Analysis of the ENCOURAGE Trial

Joshua S. Richman, PhD, MD[1,2]
Susan Andreae, MPH[3]
Monika M. Safford, MD[3]

[1] Birmingham VA Medical Center, Birmingham, Alabama

[2] Department of Surgery, School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama

[3] Department of Medicine, School of Medicine, University of Alabama at Birmingham, Birmingham, Alabama

ABSTRACT

PURPOSE Peer support intervention trials are typically conducted in community-based settings and provide generalizable results. The logistic challenges of community-based trials often result in unplanned temporal imbalances in recruitment and follow-up. When imbalances are present, as in the ENCOURAGE trial, appropriate statistical methods must be used to account for these imbalances. We present the design, conduct, and analysis of the ENCOURAGE trial as a case study of a cluster-randomized, community-based, peer-coaching intervention.

METHODS Preliminary data analysis included examination of study data for imbalances in participant characteristics at baseline, the presence of both secular and seasonal trends in outcome measures, and imbalances in time from baseline to follow-up. Additional examination suggested the presence of nonlinear trends in the intervention effect. The final analyses adjusted for all identified imbalances with accounting for community clustering by supplementing linear mixed effect models with generalized additive mixed models (GAMM) to examine nonlinear trends.

RESULTS Largely due to the location of participants across a considerable geographic area, temporal imbalances were discovered in recruitment, baseline, and follow-up data collection, along with evidence for both secular and seasonal trends in study outcome measures. Using the standard analytical approach, ENCOURAGE appeared to be a null trial. After incorporating adjustment for these temporal imbalances, linear regression analyses still showed no intervention effect. Upon further analyses using GAMM to consider nonlinear intervention trends, we observed intervention effects that were both significant ($P < .05$) and nonlinear.

DISCUSSION In community-based trials, recruitment and follow-up may not occur as planned, and complex temporal imbalance may greatly influence the analysis. Real-world trials should use careful logistic planning and monitoring to avoid temporal imbalance. If imbalance is unavoidable, sophisticated statistical methods may nevertheless extract useful information, although the potential problem of residual confounding due to other unmeasured imbalances must be considered.

Ann Fam Med 2015;13(Suppl_1):S66-S72. doi: 10.1370/afm.1790.

INTRODUCTION

Peer support interventions are pragmatic trials typically conducted in community-based settings and typically prioritizing reach and generalizability. Despite their ubiquity, neither departures from protocol nor their analytical consequences are often explicitly discussed.

In this paper we present the design, conduct, and analysis of ENCOURAGE: Evaluating Community Peer Advisors and Diabetes Outcomes in Rural Alabama (the ENCOURAGE trial) as a case study of a cluster-randomized, community-based, peer-coaching intervention. We discuss how implementation of the trial deviated from protocol and the challenges this created for study logistics and statistical analysis. We detail the evolving analytical approach, demonstrating how methods to account for unintended temporal trends can provide more information than tradi-

Conflicts of interest: authors report none.

CORRESPONDING AUTHOR

Joshua Richman, PhD, MD
1720 2nd Avenue South, KB428
Birmingham, AL 35294
jrichman@uabmc.edu

tional analyses. We conclude with recommended strategies for optimizing conduct of community-based trials that encounter unanticipated delays.

Study Design

The ENCOURAGE trial assessed whether education and volunteer community-based peer coaching together were more effective at improving diabetes outcomes than education alone. The trial was set in Alabama's Black Belt, a predominantly rural and low-income region with scarce medical resources. The design of the study has been described in detail elsewhere.[1] Main study outcomes were changes in glycated hemoglobin (HbA_{1c}), body mass index (BMI), cholesterol, blood pressure, diabetes distress, quality of life, and patient activation.

Participants were randomized at the community level to minimize contamination; because communities were geographically dispersed, recruitment and data collection took place 1 community at a time. We planned follow-up data collection for 1 year following baseline.

Seasonal and Secular Trends

As demonstrated by a previous study of a large cohort of veterans, some biometrics such as HbA_{1c} vary seasonally,[2] decreasing by 0.4% from winter to summer, a difference frequently used as the threshold for clinical relevance in study design. Yet seasonal trends are infrequently considered in analyses or discussed as confounders in this context.

More commonly recognized are secular trends: changes in population characteristics over time unrelated to the intervention. Data collection may span enough time for both seasonal and secular trends to be important.

If data collection is not balanced between study groups with respect to time, either seasonal or secular trends can be problematic. For example, if baseline HbA_{1c} data for both study groups is collected in winter but the control group is followed up the following winter while the intervention group is followed up predominantly in spring and summer, a significant effect due entirely to seasonal trends could be attributed to the intervention. Similarly, if there is a secular trend of population-wide improvement, then whichever group was followed up second may show the greater improvement solely due to the secular trend.

DIFFERENCES FROM ORIGINAL PLANS AND RESULTING ANALYTIC CHALLENGES

Recruitment and Enrollment

Plans

Enrollment and baseline data collection occurred between February and August 2010, with communities randomized using permuted blocks. The typical 2.5-hour one-way travel time to the partner communities required that we enroll 1 community at a time, with some communities requiring more than 1 date.

Reality

The original recruitment schedule had to be modified several times due to unexpected events (eg, funerals) in our partner communities, resulting in temporal imbalance in the enrollment of intervention and control participants (Figure 1, Panel A). Specifically, participants enrolled in February, July, and August were primarily in the control group while most enrollment in April, May and June was in the control group.

Analytic Challenges Created

Figure 2 shows the seasonal variation in baseline HbA_{1c} in ENCOURAGE. The imbalance in enrollment times between groups created the possibility of confounding seasonal trends and study group characteristics.

Implications

This kind of temporal imbalance at baseline can easily occur if recruitment and randomization occur in clusters, especially if logistic constraints require them to occur sequentially over a period of time-spanning seasons.

Follow-up Data Collection

Plans

We planned to follow up on participants in both groups between June 2011 and January 2012, approximately 1 year after baseline, by community, in the same order as recruitment, allowing up to 3 months deviation from this schedule.

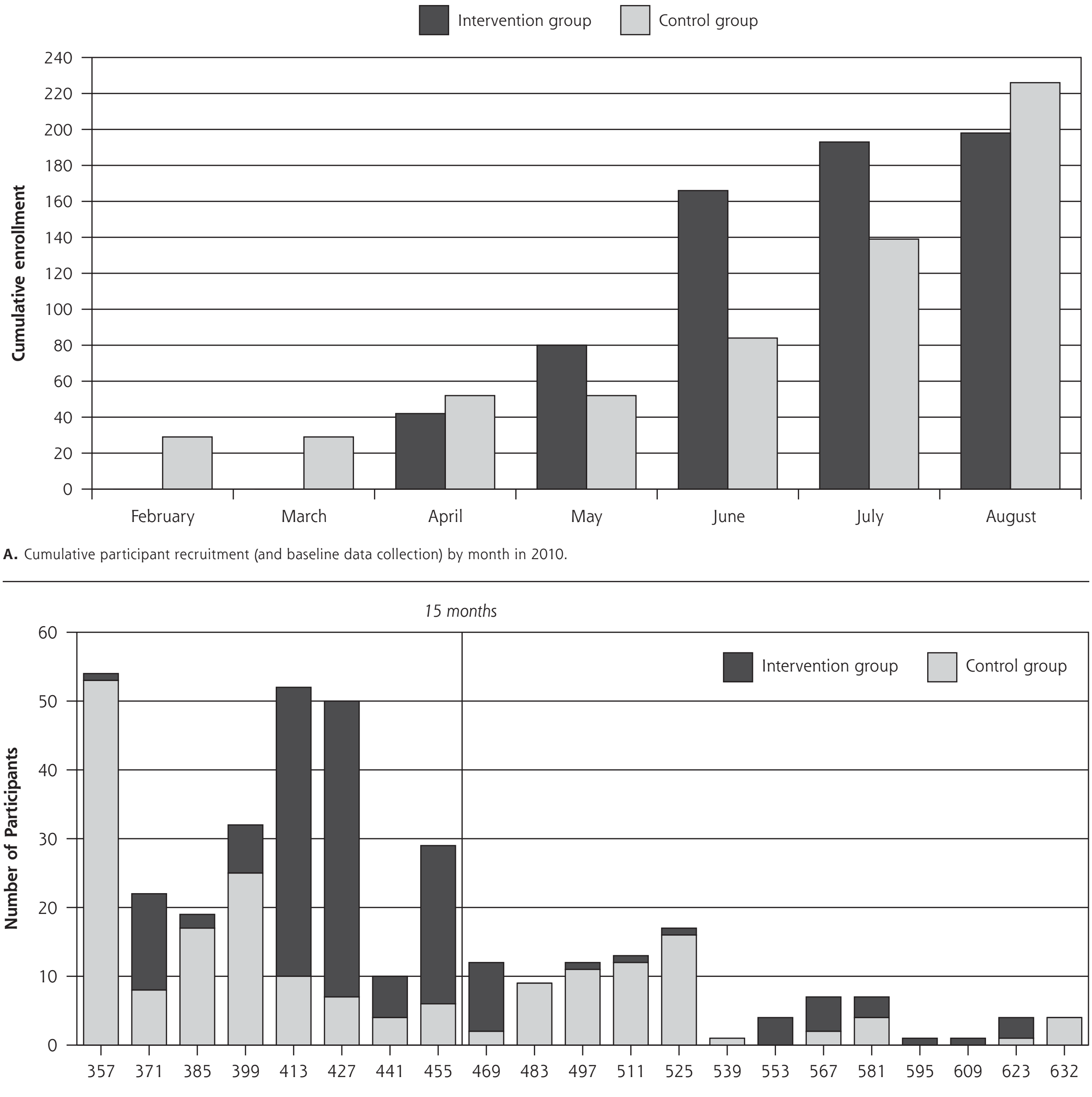
Reality

The actual course is displayed in Figure 1, Panel B. Only 268 participants (63%) had follow-up data collection within 15 months of enrollment. Although there was no significant imbalance between study groups in the proportions who had follow-up within 15 months, most of the first 50 participants followed up were in the control group.

To maximize the reach and generalizability of this pragmatic trial, we offered in-home visits for data collection, and we eventually achieved 85% retention by prolonging the follow-up period. Among participants completing follow-up more than 15 months after baseline, 62 (67%) were control and 30 (33%) were intervention participants, a considerable imbalance.

While we collected follow-up data from June 2011 to February 2012, the time elapsed from baseline to follow-up for individuals ranged from 10.3 to 20.8 months. Although there was no statistical difference in mean days to follow-up (Control 426, Intervention

Figure 1. The timing of baseline and follow-up data collection in the ENCOURAGE trial displaying temporal differences by study group.



A. Cumulative participant recruitment (and baseline data collection) by month in 2010.

B. Elapsed time from baseline data collection to follow-up in days, with a vertical line at 15 months, the standard cut-off point for 1-year follow-up. By the 15-month point, 268 participants had been followed up, of whom 51.4% (138) were in the intervention group and 48.5% (130) in the control group. After that point, 92 participants were followed up, of whom 32.6% (30) were in the intervention group and 67.4% (62) were in the control group.

436; *t*-test $P = .12$) there was a significant difference in distributions (Wilcoxon $P < .001$).

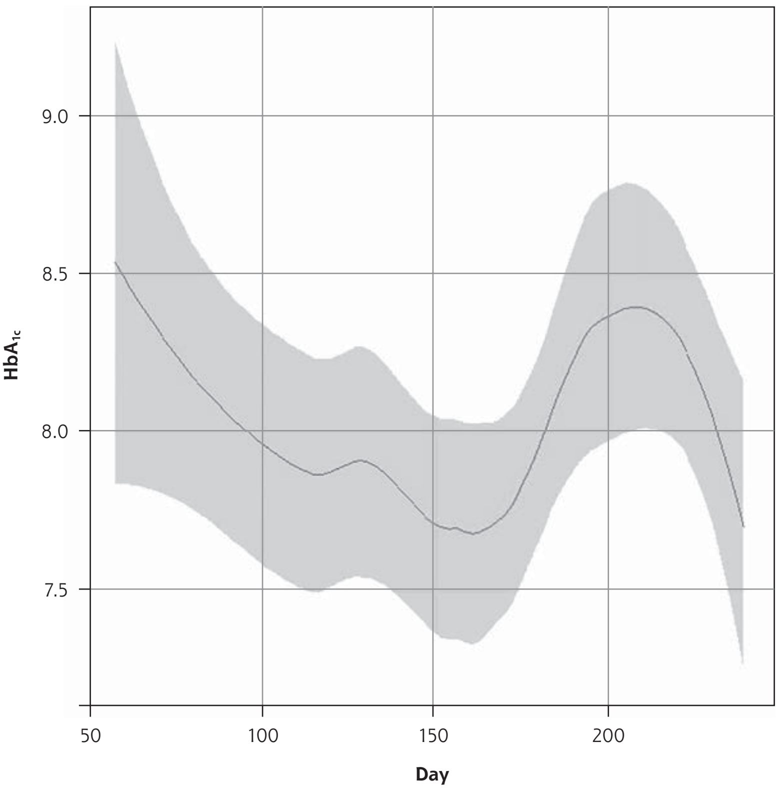
Analytic Challenges Created

Any observed effect of the intervention may have changed over time; for instance, if the peak intervention effect occurred at or before 15 months and diminished thereafter, we would expect to observe a smaller effect among participants whose follow-up data collection was much beyond 15 months. The usual approach to analyzing trial data would not reflect these temporal nuances.

Implications

The lesson for similar trials is that there can be a trade-off between adhering to protocol despite a

Figure 2. The seasonal trend in HbA_{1c} levels in the ENCOURAGE trial.



Note: The figure reflects all baseline observations reported in calendar days where January 1 is day 1 and December 31 is day 365. The period represented runs from late February to late August.

low follow-up on one hand, and breaking protocol to achieve more comprehensive follow-up with a more representative cohort on the other. The former allows for a standard analysis, while the latter requires a more complicated analysis to account for variable times to follow-up. A low follow-up (like the 63% we managed within 15 months) may bias the results of the trial, while a lengthier follow-up period may allow not just more complete follow-up, but also the possibility of tracking of changes in intervention effect over time.

Statistical Analysis

Plans

The standard analytical approach for randomized trials relies on randomization to make control and intervention study participants comparable, with a balanced time from baseline to follow-up and variability small enough to render seasonal and secular trends unimportant. The overall effectiveness of the study is estimated as a distribution-appropriate difference between groups—for example, a difference in means, change, proportion experiencing an event, or survival time. Statistical significance is tested using an unadjusted test for a difference between groups. If there are statistically significant differences between groups at baseline, the main test may take the form of regression analysis with adjustment for unbalanced characteristics. Group randomized trials typically account for clustering by using regression with generalized estimating equations or mixed effects models. Furthermore, when the primary outcomes are change in baseline characteristics such as BMI or HbA_{1c}, it is also appropriate to adjust for the individual's baseline value to control for regression to the mean.

Reality

We observed statistically significant baseline imbalances between groups in race (P <.001) and education (P = .05) with some indication of imbalance in income (P = .10). We therefore used regression models controlling for these factors and the baseline value for each outcome. For instance, our initial proposed model for change in BMI was this:

$$(1)\ \Delta \text{BMI} = \beta_0 + \beta_1\text{Group} + \beta_2\text{BMI}_{\text{baseline}} + \beta_3\text{Race} + \beta_4\text{Education} + \beta_5\text{Income}$$

To account for the challenges discussed above, we augmented this model in a stepwise fashion. First, we added a term for elapsed study time at follow-up and an interaction term allowing the groups' changes to differ over time. The new model for change in BMI was:

$$(2)\ \Delta \text{BMI} = \beta_0 + \beta_1\text{Group} + \beta_2\text{BMI}_{\text{baseline}} + \beta_3\text{Race} + \beta_4\text{Education} + \beta_5\text{Income} + \beta_6\text{Time} + \beta_7\text{Group} * \text{Time}$$

In this model the test for an intervention effect would be the significance of the group-by-time interaction term ($\beta_7 \neq 0$). An advantage of this approach is that it explicitly makes use of the information on variable exposure to the intervention. The disadvantage is that it does not estimate a single intervention effect, but rather, indicates that the difference between the groups changed significantly over time.

The standard analysis can report results such as, "We observed a decrease in mean HbA_{1c} of xx at 1 year in the intervention compared with the control group (P = .0x)." In contrast, inclusion of a group-by-time interaction changes the interpretation to "we observed that the difference in mean HbA_{1c} between intervention and control groups changed significantly over time from 1 year to 2 years (P = .0x)" The latter can't answer the question "how big was the intervention effect?" because the observed intervention effect changed over time. While this approach is valid and actually provides more information than estimating the difference at a single point in time, it is more difficult to comprehend and explain. For example, if an intervention group did worse initially but better later on, there may be no overall effect, but analysis with a group-by-time interaction might correctly identify that the differences between groups changed over time.

Second, we noted secular trends. For instance, both groups experienced an increase in LDL cholesterol, leading us to add another variable reflecting calendar time throughout the study as days from January 1, 2010. To account for seasonal variations in biometric measures, we added variables controlling for the season of baseline and follow-up data collection, defined in three-month intervals as Winter (January-March), Spring (April-June), Summer (July-September), and Fall (October-December).

The final linear model was operationalized as:

$$(3)\ \Delta\text{BMI} = \beta_0 + 1\,\beta_1\text{Group} + \beta_2\text{BMI}_{\text{baseline}} + \beta_3\text{Race} + \beta_4\text{Education} + \beta_5\text{Income} + \beta_6\text{Time} + \beta_7\text{Group} * \text{Time} + \beta_{8\text{-}10}\text{Season}_{\text{baseline}} + \beta_{11\text{-}13}\text{Season}_{\text{follow-up}} + \beta_{14}\text{CalendarTime}$$

Initial results included negative estimates for the group-by-time interactions, meaning intervention effects were decreasing over time. As an intervention takes effect, we would expect the groups to diverge, with more improvement among the intervention group, but after the intervention ends or maximum effectiveness is reached, we would expect the intervention effect to plateau or decay. The ENCOURAGE intervention was planned to last 1 year, and given that follow-up extended well beyond 1 year after baseline, we surmised that follow-up observations might have spanned both the periods of increasing effect and decay, meaning that a linear fit would be inadequate.

There are several approaches to address nonlinear trends in data. One is to add higher powers of Time to the model (eg, Time^2 or Time^3), essentially modeling the time trend as a polynomial function. While this is fairly intuitive, and polynomials readily capture effects that both rise and fall, polynomials often introduce artifacts toward the margins of the distribution (early or late follow-up times) because the higher powers dominate.[3] Another approach would be to incorporate other transformations, such as log(Time). While sometimes useful, this is less intuitive and has the potential issue that because it only increases with time, it can be less useful for modeling an effect that rises and falls. Using regression splines is a strategy that retains the advantages of polynomials while limiting the disadvantages. In essence, splines model separate sections of the data using polynomials (eg, 0 to 150 days and 150 to 300 days) but with mathematical constraints that force them to join together smoothly. Further constraints are used at the margins to reduce distortions. The points defining the intervals for separate spline fits are called 'knots' (the knots in the above example would be at 0, 150 and 300 days) and the most commonly used splines are cubic. One potential disadvantage of splines is that they add another layer of complexity and often require the user to specify the order of the splines and the location of the knots.

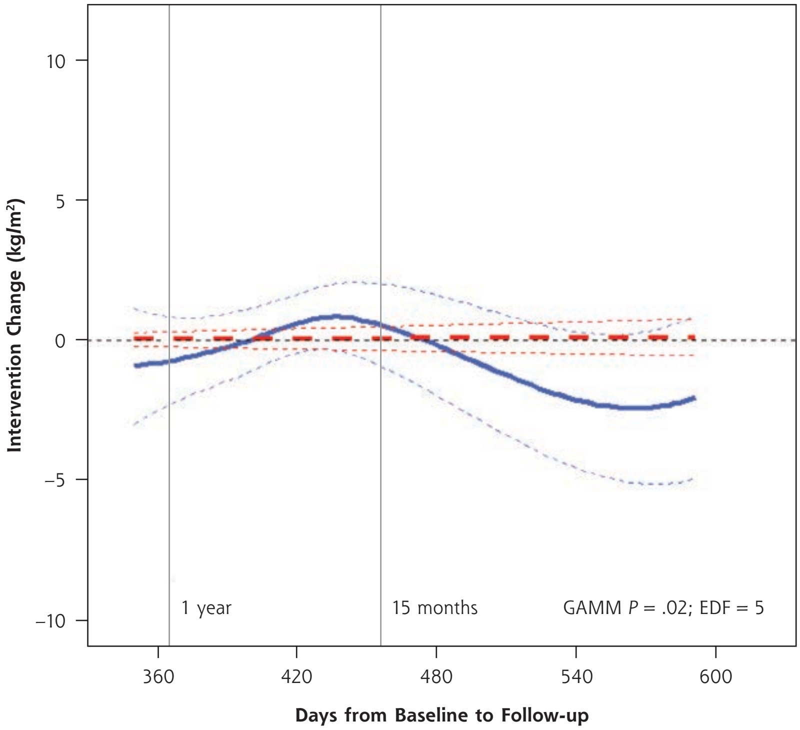
To model our temporal trends, we used generalized additive mixed models (GAMM) to consider the potentially nonlinear changing pattern of the intervention effect over time.[3,4] The fitting process iteratively optimizes the model's performance. The model is fit leaving some observations out and then validated using the excluded observations to quantify model performance. This method provides a good fit to the data without forcing the model to be linear, while penalized smoothing effectively guards against over-fitting. In our case the algorithm used splines to capture the nonlinearity, but the location and number of knots was determined iteratively by the algorithm rather than chosen explicitly. As with linear models using a group-by-time interaction, nonlinear models will not provide a single estimate of the intervention effect. Rather, they test for whether the control and intervention groups differ over time. A potential advantage is that by faithfully modeling changes over time, we could isolate any increasing intervention effect followed by plateau or decay.

When constructing the final model for each outcome, the overall time trend was fit as a linear term and the intervention effect (difference between groups) over time was allowed to be nonlinear. Initial models allowed the overall time trends to be nonlinear, but model fits provided no evidence suggesting that nonlinearity was required. The final model was similar to model (3) replacing the group-by-time terms with the penalized spline term for the intervention effect. We used an implementation of GAMM models by the package mgcv in the statistical programming language R, version 3.01, which allowed us to include community clusters as random effects.[4-6] Modeling was repeated, including the additional clustering of intervention participants within peer-advisors as well as communities, but results were not substantially different.

The results of the group-by-time models and the final GAMM fits for changes in BMI are shown in

Figure 3. The heavy dashed line shows the group-by-time linear model. The heavy solid line, in contrast, shows the statistically significant nonlinear fit of the intervention effect over time with estimated degrees of freedom (EDF) of 5 indicating the importance of the nonlinearity. Examination of the plot shows how the estimated intervention effect changed over time. Intervention group participants followed up earlier and later than 15 months experienced greater weight loss than control participants followed at the same times, but those followed around 15 months may have had slight weight gain compared with control participants. From this observation it follows that the planned time of tightly scheduled follow-up could determine whether the intervention effect was judged to be positive, negative, or non-existent.

Figure 3. The change in BMI attributable to the intervention over time, by linear model and GAMM.



Note: The plot shows the adjusted model results for the difference between intervention and control groups (the intervention effect) with 95% confidence regions bounded by thin dashed lines. The heavy dashed line shows no linear effect over time, while the heavy solid line shows the GAMM results, which allow the effect over time to be nonlinear. The GAMM $P < .05$ and the estimated degrees of freedom (EDF) >1 indicate that the association is significant and that the nonlinearity is important.

GAMM = generalized additive mixed models.

Implications

The presence of temporal imbalance requires a thoughtful look at the data to consider which imbalances may require careful examination. In the case of the ENCOURAGE trial, considering all the factors seemed to be necessary; but any given trial may or may not have evidence for secular trends, seasonal trends, or nonlinear temporal trends in intervention effect. Ultimately, we all hope to have our studies go according to plan. In the event that they do not, these approaches show a way to uncover useful information even when a simple estimate of the trial's effect is not feasible.

DISCUSSION

We demonstrate statistical approaches to account for differences between control and intervention groups in the timing of enrollment and follow-up, with the sometimes unsettling realization that the interpretation of a study's results can differ substantially depending on the analytical method. Although nonlinear analysis makes interpreting the overall study findings less straightforward because there is no single intervention effect, it offers the potential advantage of identifying the trajectory of the intervention, how long it may take to reach maximal effectiveness, and when any effect begins to diminish. This is in itself valuable information that is usually unavailable when planning trials, possibly leading to the choice of suboptimal follow-up times. The 2 approaches are not mutually exclusive; the main analysis may proceed as planned, with these more complex methods employed as secondary analyses to elucidate the main results.

Limitations remain, primarily related to confounding. It is tempting to consider the results of this study as identifying the trajectory of the intervention's effect on the study outcomes. If the follow-up times were randomly apportioned between the groups, this would indeed be the case. If the characteristics of participants with late follow-up differ from those with earlier follow-up, however, observed differences between study groups over time may be due to these unmeasured differences rather than to the trajectory of the intervention's effectiveness. If, for example, intervention group participants who lost the most weight tended to have the latest follow-up then the increased long-term weight loss observed would be due to the characteristics of the patients, not to the intervention per se. We did, in fact, observe some differences in characteristics by time to follow-up, but controlling for these in secondary analyses did not appreciably alter model results.[1]

We demonstrated that some departures from a study's planned timeline can introduce temporal imbalances and that their consequences should be carefully examined to discern whether they need to be accounted for. If so, careful and flexible statistical modeling can deal with multiple temporal factors and provide a more complete picture of the intervention effect than standard linear analysis. This requires careful analytical attention to seasonal trends, secular trends, and differential exposure to the intervention. Making the best use of the data, however, may require acknowledging that the study, as executed, precludes providing a simple and meaningful estimate of the trial's effectiveness. Furthermore, no matter how good the modeling approach, we recognize that residual confounding may remain due to temporal imbalances in recruitment and follow-up. We therefore recommend that explicit attention be paid to maintain temporal balance as closely as possible throughout future studies.

To read or post commentaries in response to this article, see it online at http://www.annfammed.org/content/13/Suppl_1/S66.

Key words: statistical data interpretation; nonlinear models; community-based trials

Submitted July 26, 2014; submitted, revised, March 2, 2015; accepted March 12, 2015.

Funding support: Funding for this research was provided by the American Academy of Family Physicians Foundation through the Peers for Progress program with support from the Eli Lilly and Company Foundation.

Previous presentation: Richman J, Andreae S, Safford MM. Temporal Trends and the Analysis of Community-Based Trials. 35th Annual Meeting of the Society of Behavioral Medicine; April 23-26, 2014; Philadelphia, Pennsylvania.

REFERENCES

1. Safford MM, Andreae SJ, Cherrington A, et al. Peer coaches to improve diabetes outcomes in rural Alabama: a cluster randomized trial. *Ann Fam Med.* 2015;13(Suppl_1):S18-S26.
2. Tseng CL, Brimacombe M, Xie M, et al. Seasonal patterns in monthly hemoglobin A_{1c} values. *Am J Epidemiol.* 2005;161(6):565-574.
3. Harrell FE. *Regression Modeling Strategies: With Applications to Linear Models, Logistic Regression, and Survival Analysis.* New York, NY: Springer; 2001.
4. Wood SN. *Generalized Additive Models: An Introduction with R.* 1st ed. Boca Raton, FL: Taylor & Francis Group, LLC; 2006.
5. R Core Team. *R: A Language and Environment for Statistical computing.* Vienna, Austria. http://www.R-project.org.
6. Wood SN. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J R Stat Soc, B.* 2011;73(1):3-36.