HHS Public Access

Author manuscript

Clin Trials. Author manuscript; available in PMC 2017 October 01.

Published in final edited form as:

Clin Trials. 2016 October ; 13(5): 504–512. doi:10.1177/1740774516646578.

Statistical Lessons Learned for Designing Cluster Randomized Pragmatic Clinical Trials from the NIH Health Care Systems Collaboratory Biostatistics and Design Core

Andrea J Cook[1,2], **Elizabeth Delong**[3,4], **David M Murray**[5], **William M Vollmer**[6], and **Patrick J Heagerty**[2]

[1]Biostatistics Unit, Group Health Research Institute, Seattle, WA USA

[2]Department of Biostatistics, University of Washington, Seattle, WA, USA

[3]Department of Biostatistics and Bioinformatics, Duke University School of Medicine, Durham, NC USA

[4]Duke Clinical Research Institute, Durham, NC USA

[5]Office of Disease Prevention, Division of Program Coordination Planning and Strategic Initiatives, Office of the Director, National Institutes of Health, Bethesda, MD USA

[6]Kaiser Permanente Center for Health Research, Portland, OR USA

Abstract

Background/Aims—Pragmatic clinical trials embedded within health care systems provide an important opportunity to evaluate new interventions and treatments. Networks have recently been developed to support practical and efficient studies. Pragmatic trials will lead to improvements in how we deliver health care and promise to more rapidly translate research findings into practice.

Methods—The NIH Health Care Systems Collaboratory was formed to conduct pragmatic clinical trials and to cultivate collaboration across research areas and disciplines to develop best practices for future studies. Through a two-stage grant process including a pilot phase (UH2) and a main trial phase (UH3), investigators across the Collaboratory had the opportunity to work together to improve all aspects of these trials before they were launched, and to address new issues that arose during implementation. Seven Cores were created to address the various considerations, including Electronic Health Records; Phenotypes, Data Standards, and Data Quality; Biostatistics and Design Core; Patient Reported Outcomes; Health Care Systems Interactions; Regulatory/Ethics; and Stakeholder Engagement. The goal of this paper is to summarize the Biostatistics and Design Core's lessons learned during the initial pilot phase with 7 pragmatic clinical trials conducted between 2012 and 2014.

Results—Methodological issues arose from the five cluster randomized trials, also called group-randomized trials, including consideration of cross-over and stepped wedge designs. We outlined general themes, challenges, and proposed solutions from the pilot phase including topics such as study design, unit of randomization, sample size, and statistical analysis. Our findings are applicable to other pragmatic clinical trials conducted within health care systems.

Conclusions—Pragmatic clinical trials using the UH2/UH3 funding mechanism provide an opportunity to ensure that all relevant design issues have been fully considered in order to reliably

and efficiently evaluate new interventions and treatments. The integrity and generalizability of trial results can only be ensured if rigorous designs and appropriate analysis choices are an essential part of their research protocols.

Keywords

Pragmatic clinical trials; cluster randomized; group randomized; electronic health record; NIH Collaboratory

Introduction

Traditional randomized clinical trials tend to be very expensive and slow to deliver results that can be implemented directly into practice.[1] On average it takes 17 years before research findings lead to widespread changes in care.[2] Randomized clinical trials tend to be conducted in a controlled environment among a carefully selected study population under ideal conditions to assess the efficacy of an intervention or treatment.[3] When actually implemented into everyday clinical practice there is often a dramatic decrease in the effectiveness. As a result, there is a need to conduct research in real world settings to provide evidence for real world practice.[4–6]

Standard practice in traditional randomized clinical trials has also led to a serious evidence paradox. There are more than 18,000 randomized clinical trials published each year along with tens of thousands of other clinical studies.[7] However, in systematic reviews we consistently report not having enough evidence to effectively inform clinical decisions for providers and patients.[4] In addition, health care delivery interventions are generally either implemented without testing, or testing is performed locally under the rubric of quality improvement. An alternative is to develop a learning health care system[8,9] that is able to provide sufficient evidence to inform clinical decisions for providers and patients.

Changing traditional practice requires the researcher to take a more practical approach to all aspects of the research design. This has led to the development of a pragmatic clinical trial (PCT) paradigm in which we more flexibly and feasibly design studies by incorporating pragmatic features into the trial.[10] There are numerous definitions for what is a pragmatic clinical trial,[4,6,10,11] but one common theme is that a pragmatic clinical trial is “designed to test intervention in the full spectrum of everyday clinical settings in order to maximize applicability and generalizability. The research under investigation is *whether an intervention actually works in real life*”.[4] A key feature in health care pragmatic clinical trials is leveraging the availability of existing data such as the electronic health record and repurposing the data for research. Another common feature has been the use of cluster, or group, randomization[12,13] for which clinics or physicians are the unit of randomization to avoid “contamination” and to allow the intervention to be applied as it would be in practice. These features present statistical challenges in the design, conduct, and analysis of pragmatic clinical trials.

This paper will present challenges encountered along with solutions developed through our experience as part of the Biostatistics and Design Core in the National Institute of Health (NIH) Health Care Systems Collaboratory (NIH Collaboratory) (https://

www.nihcollaboratory.org). One of the NIH Collaboratory's goals was to improve the way pragmatic clinical trials are conducted and to build infrastructure for collaborative research. Our goal in this manuscript is to disseminate lessons learned during the pilot phase of five cluster-randomized pragmatic clinical trials. Key topics include study design, randomization, and statistical analysis. We conclude with a discussion of unresolved issues and suggested next steps.

General study design issues

The first round of the pilot UH2 studies from the NIH Collaboratory funded seven pragmatic clinical trials in which five can be described as a variant of a cluster randomized design.[12,13] Although individually randomized trials are statistically more straightforward, cluster randomized trials are preferred when randomizing at the cluster-level facilitates the implementation of the trial, or where there is risk of contamination. For example, contamination occurs when the same provider is treating both an intervention and a control subject and (usually unconsciously) allows the treatment of one to influence the treatment of the other. Because of this "leakage" between interventions, the observed intervention effect will be diluted and biased toward the null. The studies discussed here varied in the unit of randomization and the type of cluster randomized design. We will describe some common themes across studies as well as key features that caused statistical complications which we attempted to address through design or analysis.

NIH Collaboratory motivating trial examples

We will motivate this paper through two real examples from the first round of the NIH Collaboratory. The first trial is a multi-site cluster randomized pragmatic clinical trial assessing the effectiveness of automated strategies to raise colorectal cancer (CRC) screening rates called, Strategies and Opportunities to STOP Colon Cancer in Priority Populations (STOP CRC).[14] STOP CRC randomized 26 safety net clinics within 8 health care systems to either: 1) usual care or 2) an automated, data-driven, electronic health record embedded program for mailing fecal immunochemical test (FIT) kits to patients due for colorectal cancer screening. Due to the intervention being embedded within the electronic health care record it required the intervention to be implemented at the clinic level for feasibility. The primary outcome is a clinic-level outcome defined as the proportion of patients eligible for colorectal cancer screening who complete a guaiac fecal occult blood test or fecal immunochemical test within 12 months.

The second trial is a multi-site stepped wedge cluster randomized pragmatic trial assessing the effectiveness of incorporating age- and modality-appropriate epidemiological benchmarks for common imaging findings on standard lumbar spine imaging reports to reduce spine-related intervention intensity called, Lumbar Imaging with Reporting of Epidemiology (LIRE).[15] LIRE randomized 100 primary care clinics within 4 health care systems using a stepped wedge design.[16] At period 0, or baseline, all clinics were receiving the standard imaging report (control intervention). Then clinics were randomized to the timing of when they would start receiving the new reports with epidemiologic benchmarks at 5 potential turn-on times. By the end of the trial after completion of the 5 turn-on times all

clinics will be receiving the reports with the additional epidemiological benchmark information. The primary outcome of interest is a patient level spine-related intervention intensity measured by relative value units after one year of receiving the imaging report.

Choice and number of randomization units

Cluster randomized trials embedded within health care systems have varying types of randomization units available. Specifically, one can think of a hierarchy in which a patient sees a provider, who is part of a panel, within a clinic, which is part of region, and that region is within a health care delivery system or site (Figure 1). In an ideal study design setting (no dashed lines), we have complete nesting within each organization level and there is very little potential for contamination among units at the same level. Therefore, if the study randomized one provider to one intervention and another provider to another intervention, the provider and their patients would not be exposed (i.e. contaminated) to the other intervention. We use provider as a general term which might be a primary care physician if only primary care physicians deliver the intervention, but could also mean nurse, physician assistant, specialist (e.g. radiologist) or a combination of different medical professionals depending on the type of intervention being evaluated.

However, often there is potential for contamination (dashed lines). For example, if a patient sees more than one provider in a panel of providers then there may be potential for contamination if the intervention is randomized at the provider-level (Figure 1: patient B sees both providers 2 and 3 within panel 2). To prevent contamination, panel-level randomization may be feasible. However, sometimes providers go to different clinics, or provide care across panels within the same clinic (Figure 1: provider 1 provides care at two panels). One might decide to exclude those providers from randomization or randomize at a higher level like the clinic.

Understanding the health care system is crucial to providing statistical advice on the study design. For the statistician, the goal is to work with the study team to determine the randomization unit that yields the most clusters, while still being feasible to deliver the intervention with minimal risk of contamination. With a fixed total number of patients, maximizing the number of units of randomization increases power and reduces potential bias due to imbalance of baseline cluster characteristics. Hence it is important to select the lowest level cluster that will have minimal risk of contamination. The trade-off between decreasing the risk of contamination by decreasing the number of randomization units or increasing the number of randomization units to increase power and balance is always a consideration. Further, having more clusters provides better statistical properties when analyzing cluster-level data. If the number of clusters is small (under 40 or 50), one may want to apply a small sample correction when using popular statistical approaches that handle correlated data such as generalized estimating equations.[17–20] Later we will detail ways to control for potential imbalance in randomization for situations that involve a relatively small number of clusters.

Our experience in the NIH Collaboratory was that often investigators started with a small number of large clusters, typically at the clinic-level, but through discussions with the Biostatistics and Design Core discovered that the panel or provider-level was actually feasible. This allowed the projects to increase the sample size of clusters thereby increasing

the power for assessing the same effect size or providing the same power for a smaller effect size.

Unequal cluster sizes

Another important consideration in cluster randomized designs in health care systems is that often the cluster size is variable across clusters. Substantial historical cluster design work came from the education literature in which cluster size was relatively homogenous. Classrooms have about the same number of students and therefore the effect of unequal cluster sizes was not as much of an issue. However, this is definitely not the case for cluster randomized studies within health care systems. One study in the NIH Collaboratory randomized clinics within four health care systems; one system had 11 clinics with an average of about 1400 patients per clinic, while another system had 89 clinics with an average of 72 patients per clinic. Having unequal cluster size decreases the power of the study relative to a balanced cluster design.[21] It also has issues in terms of analyses and determining which effect estimate (person-level versus cluster-level) is of most interest to the study. We will detail implications of unequal cluster size to both the sample size and the analysis approach.

Unequal cluster size and sample size—A common approach to determining sample size in cluster randomized designs is the use of a design effect which takes into account the correlation amongst observations within the same cluster. First one calculates the sample size needed if one was doing a randomized clinical trial without clustering and then this is inflated by the design effect. For a simple clustered randomized design with balanced cluster sizes the design effect (DEFF) is,

$$\mathrm{DEFF}=1+(\mathrm{n}-1)\rho$$

where ρ is the intraclass correlation coefficient (ICC) and 3 is the size of a cluster. However, in situations with a small number of clusters it is recommended to further account for the degrees of freedom that will be available for the test of the intervention effect.[22] The degrees of freedom for two group cluster randomized trials is the number of clusters minus one (degrees of freedom=n-1) and then the reference distribution of the test statistic is a t-distribution with the specified degrees of freedom. Failure to account for either of these issues will result in an inflated type I error rate.[23]

This traditional approach to determining sample size assumes that the clusters are homogenous with respect to size. However, if cluster size is variable the approach to sample size needs to be modified.[24] Eldridge et al (2009)[21] provides a summary of approaches to calculate the design effect and interclass correlation coefficient dependent on outcome type and analysis method. Note that the resulting sample of individuals needed to achieve the same power will be larger when taking into account variable cluster sizes compared to assuming a simple balanced cluster design.[21] Therefore assuming a balanced cluster design for sample size calculations will yield underpowered pragmatic clinical trials if in actuality there will be a large variability in cluster sizes.

One advantage of conducting pragmatic clinical trials within health care systems is that it is possible to actually estimate the interclass correlation coefficient and cluster sizes based on available electronic health record data. This was an important part of the UH2 pilot phase process in the NIH Collaboratory to obtain accurate estimates of both of these quantities to assure the UH3 main trials would have adequate power.

Unequal cluster size and statistical analysis—When conducting a cluster randomized trial one has choices in the target population of interest to drive the statistical estimation of the effectiveness of an intervention. For example, if the randomization unit is at the clinic-level, one could decide to a) compare the intervention effect across clinics (marginal clinic-level effect), b) compare within clinic intervention effect (within clinic effect), c) compare intervention effect across patients (marginal patient-level effect) or d) something in-between.

We will first focus on the concept of marginal clinic-level versus patient-level effects, and these relate to the assumed population over which estimation will average. With multilevel structure and multiple clinics we can define averages over the population of clinics where each would have equal weight, or we can define averages over the population of subjects where clinic summaries would need to be weighted by their cluster size. To illustrate the difference between these effects first assume a simple approach to estimate a clinic-level intervention effect by taking the mean outcome at each clinic c,

$$\hat{\mu}_c = \sum_{i=1}^{n_c} Y_{ci}/n_c,$$

where Y_{ci} is the outcome for patient i in clinic c and n_c is the number of patients in clinic c. Then to estimate the mean clinic-level difference in intervention effect one can simply estimate,

$$\hat{\Delta}^C = \frac{\sum_{c=1}^{N} \hat{\mu}_c X_c}{\sum_{c=1}^{N} X_c} - \frac{\sum_{c=1}^{N} \hat{\mu}_c (1-X_c)}{\sum_{c=1}^{N} (1-X_c)} = \hat{\mu}^C_{x=1} - \hat{\mu}^C_{x=0},$$

where X_c is 1 if clinic c was randomized to the intervention and 0 otherwise. Therefore this estimated mean difference weights each clinic equally yielding an average clinic-level effect. Such an estimate targets an average treatment effect (ATE) defined for the population of clinics, and parallels the simple definition of average treatment effect for individual randomized trials. This approach, which is used in the STOP CRC study, avoids entirely the problems otherwise associated with unequal cluster sizes as the data are reduced to provide a single estimate for each cluster.

However, if the investigators were interested in a patient-level effect one might calculate the following,

$$\hat{\Delta}^P = \frac{\sum_{c=1}^{N}\sum_{i=1}^{n_c} X_c Y_{ci}}{\sum_{c=1}^{N} X_c n_c} - \frac{\sum_{c=1}^{N}\sum_{i=1}^{n_c} (1-X_c) Y_{ci}}{\sum_{c=1}^{N} (1-X_c) n_c} = \hat{\mu}^P_{x=1} - \hat{\mu}^P_{x=0}.$$

In this setting the clustering is essentially a nuisance and does not directly relate to the definition of the population (patients) that is of interest for defining an average treatment effect. For estimation the definition of a patient-level effect leads to weighting information to validly represent the population of patients and therefore each patient contributes equal weight. Equivalently, this approach weights cluster-level summaries by the size of the cluster. If the cluster size is balanced, $n_c = n$, then the patient-level estimate is the same as the clinic-level estimate. However, this is not the case in the unequal sample size setting.

The most common approach to estimate marginal effects is through generalized estimating equations (GEE).[25] There are numerous ways to apply GEE, but one approach is to assume an independent working correlation structure and then choose the weights to estimate whichever level of inference is of interest and through robust standard errors to correct the variance for correlation due to the cluster randomized design. By using GEE one can also adjust for patient and higher-level baseline characteristics which may be advantageous to handle by chance covariate imbalance between intervention groups or increase power especially for continuous outcomes.[26,27] As noted above, if the number of clusters is small (under 40 or 50), one may want to apply a small sample correction.[17–20] Further discussion on the issues and potential solutions in estimating different population averages (e.g. patient versus clinic-level) when applying GEE have been previously published and describe this estimation issue in terms of informative cluster sizes.[28,29]

Another approach to estimate a marginal effect for continuous outcomes is to apply a linear mixed model[30] using patient-level outcomes. Typically when using a linear mixed model with continuous outcome data a cluster-level random effect or a random intercept model is used:

$$Y_{ci} \sim \beta_0 + \beta_x X_{ci} + b_c \text{ and } b_c \sim N(0, \sigma_b^2) \text{ for c}=1,\ldots,\text{N and i}=1,\ldots,n_c$$

where, b_c is the random effect for cluster c. Mixed model estimation using maximum likelihood is equivalent to use of weighted least squares with weights inversely proportional to the variance (Gauss-Markov). With no individual covariates it can be shown that this approach weights cluster-level means using the inverse of the degrees of freedom, or n_c/ [1+ (n_c − 1)ρ] this shows that weighting is intermediate to equal weights at the cluster-level which would result with ρ = 1, and weighting proportional to cluster size which corresponds to ρ = 0 Therefore, using a mixed model implicitly estimates a target parameter that can be viewed as intermediate to cluster and patient-level effects. The implicit level of effect estimated depends on amount of correlation within cluster (more correlation moves estimate closer to a cluster-level effect) and how the model is fit. Comparable estimates can be made using GEE and assuming an exchangeable working correlation structure instead of independence. However, the interpretation of the estimate as an in-between cluster and patient-level effect is not likely clinically meaningful to the scientific question of interest.

Therefore there should be caution when using such an approach when the interest is in either a patient or cluster level effect. There is potential to weight the estimate back to a scientifically meaningful quantity, but this is not readily viable with current software capabilities.

Another parameter of interest may be a within cluster effect instead of a marginal effect, which is directly estimable when using generalized linear mixed effect models.[30] Note for continuous outcomes, due to collapsibility, the generalized linear mixed effect model estimates can be interpreted as both within cluster and between (marginal) cluster effects. However, for binary outcomes when estimating an odds ratio this is not the case. The within cluster effect addresses questions like “What is the expected benefit if a clinic implements the new intervention relative to Usual Care?” instead of the marginal clinic-level effect “What is the benefit if all clinics in the health system changed to the new intervention relative to Usual Care?” Typically, one is most comfortable estimating this within clinic quantity if within each cluster they observe both interventions being compared. For example in a stepped wedge design every cluster observes a period of time on Usual Care and then crosses to the Intervention and therefore estimating a within cluster effect may be desirable.

Choice of the which quantity to estimate should be made based on the scientific question of interest, but statistical tradeoffs, including power, must also be considered. Further, given a particular unit of analysis, there are different analytical approaches that will present different power tradeoffs. Some of these have been explored recently for continuous outcomes,[21] but the literature addressing tradeoffs for binary or survival outcomes remains limited to specific designs.[31]

Which cluster randomized design?

So far we have focused on simple cluster randomized designs to illustrate general statistical concepts. We now delve into some newer cluster study designs that may have advantages especially in health care pragmatic clinical trials.

Simple cluster randomized design

We define a simple cluster randomized design as one in which randomization is at the cluster-level and in which the cluster remains on the same intervention throughout the course of the trial (Figure 2: Simple Cluster). Advantages are being relatively simple and may be easy to implement. A major disadvantage is that not all clusters get the intervention and that may not be viewed positively in a health care system. To get buy-in from a health care system to conduct a study, stating that only 50% of clusters will be randomized to the intervention may not gain approval. These are often integrated systems and if the intervention requires infrastructure, the health care system is reluctant to make an investment if they cannot at least eventually provide the new intervention to the entire system in a timely manner. This will be true especially if the intervention is likely to have a benefit (case tested in other systems) and there is minimal potential for harm.

A statistical and scientific disadvantage of clusters only receiving one intervention is that within cluster contrasts across treatment groups are not observed. Therefore, for both feasibility and statistical purposes other cluster randomized designs may be preferred.

Cluster with crossover randomized design

This design approach randomizes all clusters to an initial intervention and then, after a certain period of time, every cluster will switch, e.g. crossover, to the other intervention (Figure 2: Cluster with Crossover).[31] This study design is only feasible if the intervention can be turned off and on without "learning," such that residual practices are not carried over from one period to the next. This carryover effect would cause contamination across intervention arms. Solutions using wash out periods after the crossover, during which the data from the clusters are discarded, may help contamination, but are not always feasible.

A simple alternative is to have data collected from all clusters before the intervention period (baseline period) and then half of the clusters receive the intervention and data continue to be collected (intervention period) (Figure 2: Cluster with Partial Crossover from Baseline). This is the most common cluster randomized design. This design with an untreated baseline assessment period followed by a parallel cluster randomization has advantages statistically because data are now available from some units to efficiently estimate a within cluster effect without the potential for "learning" contamination, but this design still has the feasibility issue that not all clusters receive the intervention.

Another major statistical advantage is the power gain that is available when implementing a cross-over cluster design versus a simple cluster design. The magnitude of the improvement will depend on the cluster intraclass correlation coefficient.

Stepped wedge design

The stepped wedge design[16] was developed specifically to address the issue that feasibly, and often ethically, all clusters should eventually receive the intervention over the course of the study. This design randomizes the timing of when the intervention is turned on for a given cluster or set of clusters and was used for the NIH Collaboratory LIRE study. Once the intervention is turned on for a cluster the intervention remains on for the remainder of the study (Figure 2: Stepped Wedge). This can be thought of as a staggered cluster with cross-over design. Temporally spacing the intervention allows one to control for changes over time within the health care system. Health care systems are not static and other changes may be occurring outside of the scope of the study. Of course, one would like to limit these changes, but this may not always be possible. Randomizing the start time of the intervention allows time to be controlled for in the design.

The key advantages of the stepped wedge design are that all clusters receive the intervention, it is possible to control for external temporal trends, and one can make a within cluster interpretation if desired. One still needs to be careful about contamination across clusters. However, by gaining valuable within cluster cross-over data, power may be improved relative to simple cluster randomization[16,32] and therefore fewer clusters may be required.

There are unanswered methods questions when conducting stepped wedge designs such as the best way to conduct the mixed-effect analyses in terms of appropriately specifying the random effects and how to calculate power, as current software is relatively limited. Trials recently published special issues on stepped wedge designs touching on the design, analysis, reporting, and sample size calculation for stepped wedge designs.[33] Given the added complexities for this design, there is need for more statistical expertise when proposing such a design and one should be cautioned to think of the implications of the analysis choice and the longitudinal data structure the stepped wedge design implies.

Randomization

The last area of major discussion in the first pilot phase of the NIH Collaboratory focused on the best approach to implementing randomization for cluster randomized designs. Some studies had less than 50 clusters to be randomized and one (STOP CRC) was closer to 20. Under these conditions, crude randomization in which 50% of clusters are randomized to be in the intervention group and 50% to be in the control may not be optimal. Imbalanced baseline characteristics between arms are likely due to chance with a small number of clusters randomized to each arm; subgroup analyses may also be more challenging if simple randomization is used, because the arms may not be balanced with respect to the factors that would define the subgroups.

For typical health care studies we usually have a large amount of baseline data from the electronic health record about the clusters such as cluster size, demographic make-up, baseline outcome (e.g. baseline mean blood pressure by cluster), region, etc. Using this information can help inform the optimal randomization scheme to achieve balance across potential confounders.

There are several approaches to balance between cluster differences, including pair matching and stratification. Pair matching pairs clusters together that have similar baseline characteristics and then randomizes within pairs. However, it can be difficult to choose pairs and if one cluster in the pair drops out, the entire pair is lost in the analysis. Also, pair matching may cause complications in the analyses.[34–36] Another approach is stratification in which one creates strata based on a small set of predictors and balances randomization within strata. Stratification avoids the analytic issues created by pair matching and therefore may be preferred.[36,37]

An alternate approach is constrained randomization.[38] Here, one simulates a very large number of potential randomization schemes to attempt to represent the entire randomization space (with few enough clusters, the entire randomization space can be enumerated); remove duplicates; assesses each potential randomization scheme for baseline characteristic balance and restricts to those with sufficient balance according to a pre-specified metric; randomly selects from the set of “constrained” balanced randomization schemes a single randomization assignment; and randomly assigns the intervention groups to that selected scheme.

This approach seemed like a viable option to the Biostatistics and Design Core, but it was unclear whether constrained randomization was better than other randomization approaches in terms of Type 1 error and power and what would be the best way to implement the approach including implications for the analysis. We conducted a simulation study comparing constrained to crude randomization,[39] and briefly summarize that work here. For continuous and normally distributed outcomes, results indicated that the analysis should still adjust for the balanced potential confounders. In addition the adjusted F-test and the permutation test performed similarly and slightly better for constrained randomization in terms of power. However, when performing the permutation test, the constrained randomization space should be used.

Overall, there are advantages to using available information to attempt to balance randomization for important baseline characteristics. In particular cluster size is important to balance since that has direct implication on power and those with similar cluster sizes often have other characteristic similarities (e.g. large clinics are often in denser part of the city with more similar patient characteristics).

Discussion

Pragmatic clinical trials are extremely important to reflect real world settings and to move research quickly into practice. Statisticians working on these trials need to be flexible, but still assure that the findings of pragmatic clinical trials are unbiased, efficient, generalizable, and replicable. The first question needs to be, “Should this study be addressed using a pragmatic clinical trial approach?” Pilot efficacy and feasibility studies, especially in terms of proof of concept, are still needed before moving to pragmatic clinical trials which tend to be large simple studies without as much control or oversight as more traditional randomized clinical trials. Of course, patient safety is a top priority and is often why a pragmatic clinical trial design may not be feasible if monitoring of safety outcomes cannot be achieved using electronic health record data. Best practices for data safety monitoring of pragmatic clinical trials are being developed and are beyond the scope of this manuscript.

There are open statistical questions using electronic health record for data safety monitoring especially in terms of data lag issues. Different health care systems have different data lag timing and a statistician needs to understand the implications of such lagged data (e.g. if the patient goes out of network it may take months before the billing claim information arrives in the system). Therefore it may be statistically better to not use all the available data at an interim monitoring analysis, but instead incorporate the data lag time so that you have relatively complete information to assess an unbiased estimate.

Another major factor is that new information on a patient is only observed if that patient interacts with the delivery system. For example, if an intervention improves the health of a patient, that patient may discontinue interacting with the health care system, which, although the intervention was successful has implications on missing outcome information on which to evaluate the intervention. Defining outcomes appropriately and choosing the correct statistical approach requires intimately understanding the health care delivery system and implications for outcome assessment.

Health care pragmatic clinical trials provide statistical challenges, but are needed to address important “real world” questions. This paper discussed some design and analysis challenges and solutions based on actual pragmatic clinical trials being conducted within the NIH Collaboratory. As pragmatic clinical trials continue to be conducted there will certainly be numerous new statistical challenges in this very important research area.

Acknowledgments

Grant Support: National Institutes of Health grants 1U54AT007748-01, 1UH2AT007769-01, 1UH2AT007782-01, 1UH2AT007755-01, 1UH2AT007788-01, 1UH2AT007766-01, 1UH2 AT007784-01, and 1UH2AT007797-01.

The views presented here are solely the responsibility of the authors and do not necessarily represent the official views of the National Institutes of Health.

References

1. Zwarenstein M, Oxman A. Why are so few randomized trials useful, and what can we do about it? J Clin Epidemiol. 2006; 59:1125–1126. [PubMed: 17027421]
2. Morris ZS, Wooding S, Grant J. The answer is 17 years, what is the question: understanding time lags in translational research. J R Soc Med. 2011; 104:510–520. [PubMed: 22179294]
3. Weiss NS, Koepsell TD, Psaty BM. Generalizability of the results of randomized trials. Arch Intern Med. 2008; 168:133–135. [PubMed: 18227357]
4. Patsopoulos NA. A pragmatic view on pragmatic trials. Dialogues Clin Neurosci. 2011; 13:217–224. [PubMed: 21842619]
5. Treweek S, Zwarenstein M. Making trials matter: pragmatic and explanatory trials and the problem of applicability. Trials. 2009; 10:37. [PubMed: 19493350]
6. Tunis SR, Stryer DB, Clancy CM. Practical clinical trials: increasing the value of clinical research for decision making in clinical and health policy. JAMA. 2003; 290:1624–1632. [PubMed: 14506122]
7. Chalmers I, Bracken MB, Djulbegovic B, et al. How to increase value and reduce waste when research priorities are set. Lancet. 2014; 383:156–165. [PubMed: 24411644]
8. Etheredge LM. A rapid-learning health system. Health Aff. 2007; 26:107–118.
9. Greene SM, Reid RJ, Larson EB. Implementing the learning health system: from concept to action. Ann Intern Med. 2012; 157:207–210. [PubMed: 22868839]
10. Thorpe KE, Zwarenstein M, Oxman AD, et al. A pragmatic-explanatory continuum indicator summary (PRECIS): a tool to help trial designers. J Clin Epidemiol. 2009; 62:464–475. [PubMed: 19348971]
11. MacPherson H. Pragmatic clinical trials. Complement Ther Med. 2004; 12:136–140. [PubMed: 15561524]
12. Donner, A.; Klar, N. Design and analysis of cluster randomization trials in health research. London: Arnold; 2000.
13. Murray, DM. Design and analysis of group-randomized trials. New York, NY: Oxford University Press; 1998.
14. Coronado GD, Vollmer WM, Petrik A, et al. Strategies and Opportunities to STOP Colon Cancer in Priority Populations: design of a cluster-randomized pragmatic trial. Contemp Clin Trials. 2014; 38:344–349. [PubMed: 24937017]
15. Jarvik JG, Comstock BA, James KT, et al. Lumbar Imaging With Reporting Of Epidemiology (LIRE)-Protocol for a pragmatic cluster randomized trial. Contemp Clin Trials. 2015; 45:157–163. [PubMed: 26493088]
16. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. Cont Clin Trial. 2007; 28:182–191.
17. Fay MP, Graubard BI. Small-sample adjustments for Wald-type tests using sandwich estimators. Biometrics. 2001; 57:1198–1206. [PubMed: 11764261]

18. Kauermann G, Carroll RJ. A note on the efficiency of sandwich covariance matrix estimation. J Am Stat Assoc. 2001; 96:1387–1396.
19. Mancl LA, DeRouen TA. A covariance estimator for GEE with improved small-sample properties. Biometrics. 2001; 57:126–134. [PubMed: 11252587]
20. Morel JG, Bokossa MC, Neerchal NK. Small sample correction for the variance of GEE estimators. Biom J. 2003; 4:395–409.
21. Eldridge SM, Ukoumunne OC, Carlin JB. The intra-cluster correlation coefficient in cluster randomized trials: A review of definitions. Int Stat Rev. 2009; 77:378–394.
22. Cornfield J. Randomization by group: a formal analysis. Am J Epidemiol. 1978; 108:100–102. [PubMed: 707470]
23. 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; 20:313–337. [PubMed: 10182207]
24. Donner A. Sample size requirements for stratified cluster randomization designs. Stat Med. 1992; 11:743–750. [PubMed: 1594813]
25. Zeger SL, Liang KY. Longitudinal data analysis for discrete and continuous outcomes. Biometrics. 1986; 42:121–130. [PubMed: 3719049]
26. Moore KL, van der Laan MJ. Covariate adjustment in randomized trials with binary outcomes: Targeted maximum likelihood estimation. Stat Med. 2009; 28:39–64. [PubMed: 18985634]
27. Robinson LD, Jewell NP. Some surprising results about covariate adjustment in logistic regression models. Int Stat Rev. 1991; 59:227–240.
28. Huang Y, Leroux B. Informative cluster sizes for subcluster-level covariates and weighted generalized estimating equations. Biometrics. 2011; 67:843–851. [PubMed: 21281273]
29. Seaman S, Pavlou M, Copas A. Review of methods for handling confounding by cluster and informative cluster size in clustered data. Stat Med. 2014; 33:5371–5387. [PubMed: 25087978]
30. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. [PubMed: 7168798]
31. Turner RM, White IR, Croudace T. Group PIPS. Analysis of cluster randomized cross-over trial data: a comparison of methods. Stat Med. 2007; 26:274–289. [PubMed: 16538700]
32. Rhoda DA, Murray DM, Andridge RR, et al. Studies with staggered starts: multiple baseline designs and group-randomized trials. Am J Public Health. 2011; 101:2164–2169. [PubMed: 21940928]
33. [accessed 17 August 2015] Stepped wedge randomized controlled trials. 2015. http://www.trialsjournal.com/series/SteppedWedge
34. Diehr P, Martin DC, Koepsell TD, et al. Breaking the matches in a paired t-test for community interventions when the number of pairs is small. Stat Med. 1995; 14:1491–1504. [PubMed: 7481187]
35. Donner A, Taljaard M, Klar N. The merits of breaking the matches: a cautionary tale. Stat Med. 2007; 26:2036–2051. [PubMed: 16927437]
36. Imbens, GW. [accessed 14 May 2014] Experimental design for unit and cluster randomized trials. 2011. http://cyrussamii.com/wp-content/uploads/2011/06/Imbens_June_8_paper.pdf
37. Donner A, Klar N. Pitfalls of and controversies in cluster randomization trials. Am J Public Health. 2004; 94:416–422. [PubMed: 14998805]
38. Moulton LH. Covariate-based constrained randomization of group-randomized trials. Clin Trials. 2004; 1:297–305. [PubMed: 16279255]
39. Li F, Lokhnygina Y, Murray DM, et al. An evaluation of constrained randomization for the design and analysis of group-randomized trials. Stat Med. Epub ahead of print 23 November 2015.

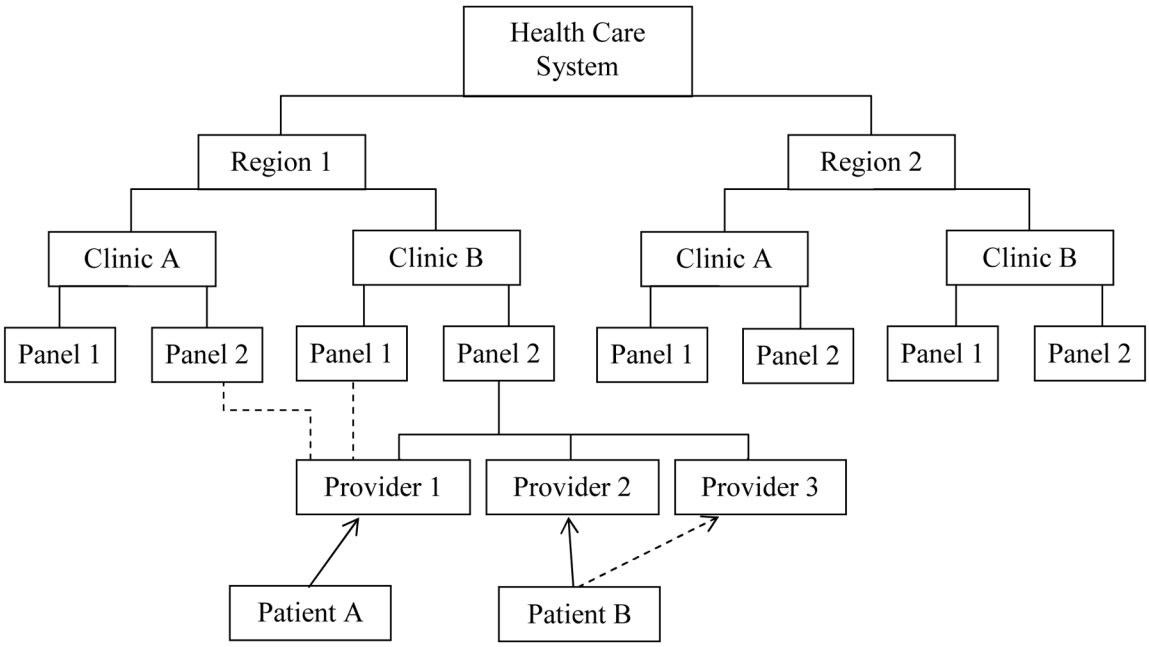

Figure 1.
Example of a common configuration of a health care system.

	Cluster	Baseline	Period 1	Period 2	Period 3	Period 4
Simple Cluster	1	--	UC	UC	UC	UC
	2	--	INT	INT	INT	INT
	3	--	INT	INT	INT	INT
	4	--	UC	UC	UC	UC
Cluster with Crossover	1	--	UC	UC	INT	INT
	2	--	INT	INT	UC	UC
	3	--	INT	INT	UC	UC
	4	--	UC	UC	INT	INT
Cluster with Partial Crossover from Baseline	1	UC	UC	UC	UC	UC
	2	UC	INT	INT	INT	INT
	3	UC	INT	INT	INT	INT
	4	UC	UC	UC	UC	UC
Stepped Wedge*	2	UC	INT	INT	INT	INT
	3	UC	UC	INT	INT	INT
	4	UC	UC	UC	INT	INT
	1	UC	UC	UC	UC	INT

* Stepped Wedge randomizes clusters to which sequence of the intervention is to be given while the other cluster designs randomize the intervention assignment to a specific cluster

Abbreviations: UC = Usual Care, INT = Intervention

Figure 2.
Different Cluster Randomized Design Configurations