Original Article

Kidney Res Clin Pract 2021;40(1):62-68
pISSN: 2211-9132 • eISSN: 2211-9140
https://doi.org/10.23876/j.krcp.20.104

KIDNEY RESEARCH AND CLINICAL PRACTICE

Comparison of chronic kidney disease trial designs and analysis strategies

John Lawrence

Office of Biostatistics, Center for Drug Evaluation and Research, U.S. Food and Drug Administration, Silver Spring, MD, United States

Background: Despite the large burden of chronic kidney disease (CKD), it is challenging to conduct adequately powered clinical trials in this setting. Sound and efficient trials are needed to advance treatment. Various analysis strategies can be used to compare the efficacy of a parallel trial design with that of three two period trial designs.

Methods: The type 1 error rates and powers of various trial designs were calculated using simulated data from models fit to two recent CKD trials. In addition, we assessed the influences of a variety of analysis strategies and of the presence of a carryover effect.

Results: The parallel and crossover designs (with analysis of change from baseline to the off treatment value) maintained the target type 1 error rate in all scenarios. In some scenarios, an open label design yielded inflated type 1 error rates. In many scenarios, the open label and delayed start designs had unacceptably low power and high type 1 error rates. Overall, the crossover design had the highest power by far, and always controlled the type 1 error rate.

Conclusion: The recommended approach to a CKD trial is a two period design with an endpoint that is the rate of change in estimated glomerular filtration rate from pretreatment to off treatment. As compared to a parallel trial, a crossover study involves a considerably smaller sample size and shorter total follow-up duration. A crossover design may also be preferable for patients, and facilitates recruitment of a sufficient number of subjects.

Keywords: Bioethics, Computing methodologies, Kidney diseases, Treatment switching

Introduction

Chronic kidney disease (CKD) affects an estimated 8% to 16% of adults worldwide [1]. Despite its wide prevalence, it has been challenging to run large and adequately powered randomized trials in patients with CKD [2]. Large, multi-year trials are expensive to conduct. Trials of patients with kidney disease often exclude a large proportion of the potential subjects.

Finally, a high level of nonadherence among enrolled subjects can sharply reduce a study's statistical power. A sound and efficient trial design that can handle missing data is essential to conducting a successful trial that advances treatment.

Received: January 15, 2020; **Revised:** October 29, 2020; **Accepted:** November 20, 2020
Editor: Hakmook Kang, Vanderbilt University, Nashville, USA
Correspondence: John Lawrence
Center for Drug Evaluation and Research, U.S. Food and Drug Administration, 10903 New Hampshire Avenue, Silver Spring, MD 20993, United States. E-mail: john.lawrence@fda.hhs.gov
ORCID: https://orcid.org/0000-0002-9892-2753

CKD progresses slowly. Therefore, definitive clinical endpoints, such as the need for dialysis or kidney transplant, have been replaced by surrogate endpoints that are related to kidney function or damage. These surrogate endpoints can be measured in a shorter follow-up time than that required for more definitive endpoints like the need for dialysis. A meta analysis of over 60,000 subjects in 47 randomized clinical trials [3] concluded that the slope of the estimated glomerular filtration rate (eGFR) is a viable surrogate for clinical endpoints in CKD trials. Indeed, the U.S. Food and Drug Administration supports the use of eGFR slope as a surrogate endpoint in trials of therapies for rare types of CKD [4].

Various approaches have been suggested to analyze data on the eGFR slope in randomized controlled trials [3]. For clinical trials, acute slope is defined as that 'from randomization to the first 3 months in follow-up,' while the chronic slope is that 'from 3 months to end of the trial.' Finally, the total slope is that 'from randomization to 1, 2, 3, or 4 years.' Many factors can influence the decision to use chronic slope versus the total slope in a clinical trial. Acute effects can complicate the interpretation of the treatment effect on both chronic and total slopes. A negative acute effect can attenuate or reverse the statistical power advantages of the total slope compared to the clinical end point. A negative acute effect can also increase the risk that use of the chronic slope as a surrogate end point could lead to a type 1 error relative to the clinical end point. Therefore, the acute, chronic and total slopes may all need to be assessed. One approach is to compare the chronic slopes by excluding the first 3 months of data from both arms, and then fitting a mixed effects model to the remaining data and testing for a difference in the mean slopes. An alternative method is to analyze all of the data with an acute effect term in the model so that the data are assumed to arise from a mixed effects model comprising two different slopes (a piecewise linear model). After this model is fit, the total slope is estimated by dividing the predicted mean change from baseline by the duration of treatment. This estimate is not technically a slope, because the mean trajectory is not a straight line but rather a rate of change. Both of these approaches assume that subjects stay on the treatments to which they were randomized for the duration of follow-up.

A different approach to estimating the rate of change (without the acute effect) is to have all of the subjects withdraw from the study drug at the end of a fixed period. Then, one can make end of trial measurements after the acute effect has worn off. This approach was successfully used in the Replicating Evidence of Preserved Renal Function: an Investigation of Tolvaptan Safety and Efficacy (REPRISE) trial in autosomal dominant polycystic kidney disease (ADPKD) [5]. The REPRISE trial was also notable for its use of two run in phases in order to minimize loss to follow up. Although some subjects discontinued the study drug during the trial, 96% of randomized subjects stayed in the trial and attended the final visit at 12 months. The subjects who withdrew from the trial were included in the analysis. Their annualized rate of change was estimated at the time of withdrawal by dividing their change in eGFR from baseline by their duration of follow-up.

In this study, the efficacy of a traditional randomized controlled trial design (parallel trial) with that of three two period trial designs was compared (Fig. 1). In each of the two period designs, the eGFR was measured at baseline, at the end of period 1, and at the end of period 2. A withdrawal phase at the conclusion of each trial period permits off drug measurement of the eGFR. The first design is the open label trial, during which all subjects receive the experimental drug in period 1, and no drug in period 2. The second design is the delayed start trial [6], during which the subjects were double blinded and randomized to either the experimental drug or placebo in period 1; during period 2, the subjects received the experimental drug on an open label basis. The third design is the crossover trial, during which subjects were double-blinded and randomized to receive both treatments (including one in period 1 and the other in period 2).

Using simulated data from models fit to two recent trials of CKD treatments, the type 1 error rates and the powers of these trial designs were calculated. The influence of these data on the results of a variety of analysis strategies and on the presence of a carryover effect were assessed. The objective of this work is to compare different designs in terms of the number of patients and total follow-up duration needed to achieve the objectives of a clinical trial under different scenarios. Wherever appropriate, recommendations are given for trial designs with their respective rationales.

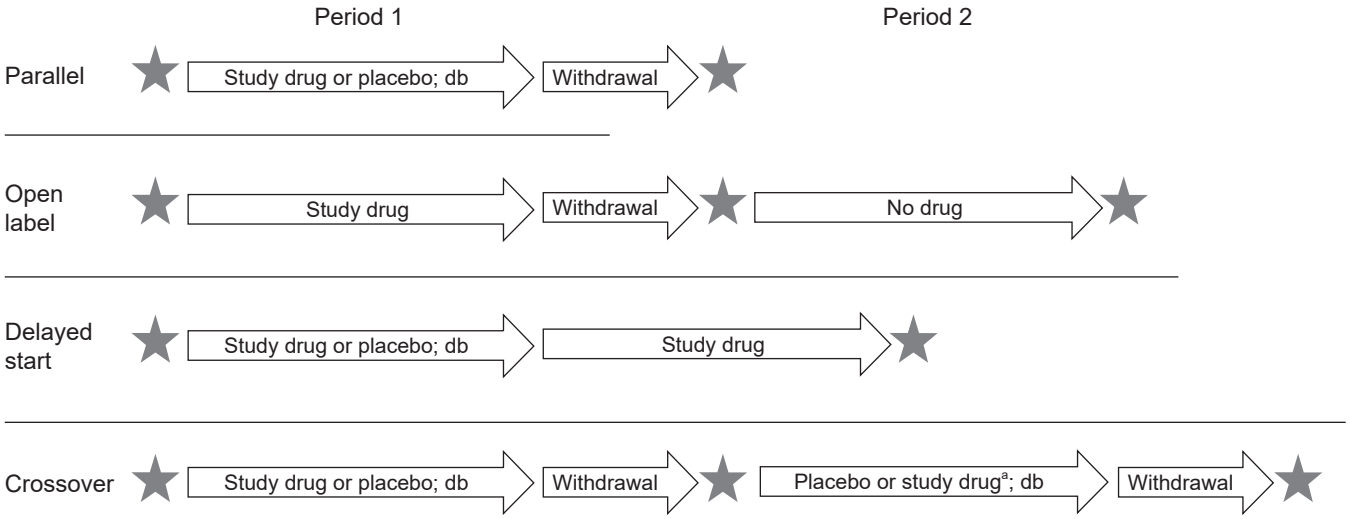

Figure 1. The parallel trial and the three designs of two period trials. Stars indicate the timepoints of estimated glomerular filtration rate measurements.
db, double blind.
[a]Whichever treatment a given subject did not receive in period 1.

Methods

Data were simulated using a mixed effects model of the following form:

$$Y_{ij} = X_i + \beta_1(t_{ij}) + \beta_2 u_{ij} + \beta_3 \times (t_{ij} - u_{ij}) I(u_{ij} > 0) + b_{1i} + b_{2i} t_{ij} + e_{ij}$$

In this model, Y_{ij} is the observed eGFR of subject i at time t_{ij}; X_i is the ideal unobserved GFR for subject i at baseline; u_{ij} is the amount of time up to time t_{ij} that the subject was on treatment; $I(u_{ij} > 0)$ is 1 if $u_{ij} > 0$, and 0 otherwise. β_1, β_2, and β_3 are fixed effects terms. $\beta_1(\bullet)$ is a function that describes the trajectory in the placebo group. When there is a constant rate of change, this term can be replaced by $\beta_1 \times t_{ij}$. β_2 is the effect of the treatment on the chronic slope and β_3 is the carryover effect; b_{1i} and b_{2i} are random effects assumed to be normally distributed with a mean of zero. In order to illustrate the difference between t_{ij} and u_{ij}, assume patient i is assigned to the treatment in period 1 and that the duration of treatment is T. Then, $u_{ij} = T$ whenever $t_{ij} > T$. The residual error terms in the model, e_{ij}, are assumed to be mutually independent and normally distributed as $N(0,\sigma^2)$; they are also assumed to be independent of the random effects. No acute effect is used in the model, because it is assumed that all of the measurements were made while the patient was off of treatment.

The data from the Tolvaptan Efficacy and Safety in Management of Autosomal Dominant Polycystic Kidney Disease and its Outcomes (TEMPO) 3-4 study [5] and REPRISE trials [7] were used to identify reasonable parameter values to simulate data from trials of each design. These trials studied patients with ADPKD. ADPKD causes bilateral, progressively enlarging kidney cysts. Despite progressive growth of the kidney cysts over a patient's lifetime, the early course of ADPKD is actually characterized by hyperfiltration and relatively normal GFR for many decades. This feature of ADPKD makes GFR an insensitive marker of underlying renal parenchymal damage. It may be necessary to consider the cause of kidney disease when determining an individual patient's response. The parameters were also modified to investigate different possible scenarios under the null hypothesis (no treatment effect) and the alternative hypothesis (beneficial treatment effect). The parameter values used are shown in Supplementary Appendix 1 (available online). The rate of change during placebo treatment was –4 mL/min per 1.73 m^2 annually, while that during experimental treatment was –3, for a chronic treatment effect of 1. The exception was in the case of a carryover effect in the crossover design (explained below).

In the TEMPO trial, subjects were randomized to receive the experimental drug or placebo, and were followed for 3 years before continuing to an open label extension. In

the REPRISE trial, the subjects were followed for 1 year. Therefore, it was assumed that the follow-up duration is 2 years per period. This follow-up is the average of the 1- and 3-year follow-up periods in the two referenced studies. For two-period designs, it is ethical and feasible for patients to consent to at least 2 years of follow-up on experimental treatment and at most 2 years on placebo.

In all designs, two eGFR measurements were assumed at baseline and at each timepoint when all subjects were off the study treatment. In the parallel and delayed-start designs, two measurements were taken at baseline and two at the end of the study. In the other two period designs, two measurements were taken at baseline, two at the end of period 1, and two at the end of period 2.

For the crossover design, two different scenarios are considered for the potential carryover effect. This design was meant to allow for the possibility that the drug imparts a structural change that persists after the drug is stopped. The first scenario was the absence of a carryover effect. In the second scenario, there was a moderately large carryover effect equal to 25% of the chronic effect in period 1. In other words, one quarter of the effect of the drug on the chronic slope was assumed to remain in period 2 for the subjects who were randomized to the experimental drug in period 1. In addition, those subjects' mean slope in period 2 (when they were not taking the drug) increased by 25% of the increase in period 1 (when they were taking the drug).

Analysis

For the parallel and delayed-start designs (Fig. 1), we averaged each subject's two baseline values and the two end of study values. A subject's change from baseline was the difference between the two average values. The annualized rate of change was the change from baseline divided by the duration of follow-up. Finally, a two sample t test is used to compare the two arms.

For the crossover design (Fig. 1), three analysis strategies were used. The first strategy was to fit a mixed effects model that included a common chronic effect for the treatments in periods 1 and 2. The null hypothesis was that the chronic treatment effect is zero. The likelihood ratio test was used to test this null hypothesis. In the other two analysis strategies, we first calculated the averages of the two measurements taken at baseline, at the end of period 1, and at the end of period 2. Next, a single value for period 1 was calculated by subtracting the baseline from the end of period 1 value, and then dividing by the duration of follow-up. A single value for period 2 was calculated by subtracting the end-of-period 1 value from the end-of-period 2 value and then dividing by the duration of follow-up. Finally, either a pooled test or two stage test was performed [8]. The two stage test started by testing for a statistically significant carryover effect that was large enough to analyze based on the data of period 1 being more powerful than are the pooled data from periods 1 and 2. If a significant carryover effect was observed, then the period 1 data were used alone as if it were a parallel trial. The significance level must be adjusted to evaluate the treatment effect in the second period (unpublished data). The preliminary test for carryover is correlated with the test of treatment effect from the first period alone. Therefore, the actual significance level of the two-stage procedure is higher than is nominal level α, even when there is no residual carryover.

For the two-period open-label design, the first averages of the two measurements taken at baseline (at the end of period 1) and at the end of period 2 were calculated. Next, a single value for period 1 was calculated by subtracting the baseline from the end of period 1 value and dividing by the duration of follow-up. A single value was calculated for period 2 by subtracting the end of period 1 value from the end-of-period 2 value and dividing by the duration of follow-up. The treatment effect for each subject was calculated by subtracting the annualized rate of change in period 2 from that in period 1. This is mathematically equivalent to the following formula: [2 × {end-of-period 1 value} – {baseline value} – {end-of-period 2 value}] ÷ [duration of each period]. The duration of each period was assumed to be equal. The numerator can also be rewritten as follows: [{end-of-period 1 value} – {baseline value} – {(end-of-period 2 value) – (end-of-period 1 value)}]. Therefore, the numerator was the difference of the treatment effect between the two periods. A one-sample t test was then performed to determine whether the mean treatment effect across subjects was greater than zero.

Asymptotic relative efficiency

When the treatment effect is small, a large sample size is needed to achieve a given power. The ratio of the sample

sizes needed is termed the asymptotic relative efficiency (ARE). The ARE was calculated to compare several trial designs and analysis strategies.

Type 1 error rate

Three different scenarios in which there was no treatment effect are considered. The natural history of subjects in a trial will depend on their characteristics, including disease stage and external factors. These factors cannot necessarily be predicted in advance or controlled by the trial's eligibility criteria. Renal function estimating equations are not linear functions of age.

Therefore, a constant rate of change cannot always be expected. In scenarios in which there is no treatment effect, $\beta_2 = \beta_3 = 0$. In the first scenario, a constant rate of change over time was assumed (decline of 4); that is $\beta_1(t_{ij}) = -4\ t_{ij}$. In the second scenario, the natural history of the rate of change was assumed to decline slightly over time. This rate of change was defined by a decline of 4.0 annually in period 1 and 3.5 annually in period 2, as follows: $\beta_1(t_{ij}) = -4\ t_{ij}$ when $t_{ij} \leq 2$ and $\beta_1(t_{ij}) = -8 - 3.5\ (t_{ij} - 2)$ when $t_{ij} > 2$. The third scenario assumed an increasing rate of change over time. This rate of change was defined by a decline of 4.0 annually in period 1 and 4.5 annually in period 2, as follows: $\beta_1(t_{ij}) = -4\ t_{ij}$ when $t_{ij} \leq 2$ and $\beta_1(t_{ij}) = -8 - 4.5\ (t_{ij} - 2)$ when $t_{ij} > 2$. The targeted type 1 error rate used for the hypothesis tests was the conventional one sided 0.025.

Power

Two different scenarios were investigated, including those with and without a carryover effect. The placebo arm was assumed to have a constant decline of 4. If there were no carryover effect, the chronic effect would be equal to 1. If there were a carryover effect, the treatment decline would be 3.75 (for a carryover effect of 0.25).

Results

Type 1 error rate

All of the designs and analyses had the target type 1 error rate, with the exception of the open label two period design with an increasing rate of change (Table 1). Therefore, when the rate of change is not constant over time (even by a small margin), there can be a marked effect on the type 1 error rate. This phenomenon is unrelated to any potential bias caused by unblinding. It is only the result of different rates of change in periods 1 and 2. It was assumed that the experimental treatment was given in period 1; however, the problem related to an inconstant rate of change can also occur if the experimental treatment is applied in period 2.

Power

Table 2 shows the observed power for the different designs and tests under various scenarios. For parallel and delayed-start designs, no patient undergoes follow-up on placebo after taking the experimental treatment. Therefore, no carryover effect is assumed. The delayed start design had the lowest power in this scenario. The crossover design had the greatest power, which did not differ considerably among the three analysis strategies. In general, a mixed effects model may be attractive, particularly when there are partial missing data. For example, if most subjects have two observations per visit but some only have one observation, a mixed effects model would handle this by assigning more weight

Table 1. Type 1 error rate (n = 500)

Design and analysis	Rate of change scenario		
	Constant	Declining	Increasing
Parallel	0.025	0.025	0.025
Open-label two-period	0.025	0.000	0.605
Delayed start	0.025	0.025	0.025
Crossover mixed effects	0.025	0.025	0.025
Crossover pooled	0.025	0.025	0.025
Crossover two-stage	0.025	0.025	0.026

100,000 simulated trials; margin of error = 0.001.

Table 2. Power by study design (n = 500)

Design and analysis	Scenario	
	No carryover effect	Carryover effect
Parallel	0.826	NA
Open-label two-period	0.993	0.916
Delayed start	0.471	NA
Crossover mixed effects	0.995	0.977
Crossover pooled	0.994	0.964
Crossover two-stage	0.989	0.963

100,000 simulated trials; margin of error ≤ 0.003.
NA, not applicable.

to the patients with two observations. The mixed effects model also appropriately handles variability in the timing of observations. For example, if period 1 ends at 2 years and one patient has two end of period 1 measurements (taken at visits on days 710 and 740), the mixed effects model uses those exact days in the model, while other analyses do not. However, the efficiency of the mixed effects model comes at the cost of assuming that the correct model is used. Other analyses do not make any strong model assumptions. The mixed effects model is a likelihood based method. Other recommended methods (that were not considered here) include multiple imputation and Bayesian approaches [9].

Asymptotic relative efficiency

The mathematical calculations for the ARE are provided in Supplementary Appendix 1. We assumed that two measurements were taken at baseline and two at the end of the study in order to assess the impact of multiple measurements at each time point in the parallel design. If only one measurement was made at each time point, then approximately 56% more subjects would be needed to achieve adequate power (than if two measurements were made at each point). The ARE was calculated at 1.56.

In comparing the crossover design with pooled analysis to the parallel design, the ARE was approximately 2.38. The gain in efficiency of the crossover design was in part a result of the additional follow-up of each subject. This gain in efficiency was also attributable to the elimination of the between subject variability in the random slope. Importantly, a crossover design involves a twofold greater duration of follow-up for each subject. Therefore, in order to achieve the same power in a trial of parallel design to one in a crossover design, the total duration of follow-up would need to be 19% greater (2.38-fold the number of subjects, each of whom was followed up for half as much time).

Discussion

The crossover design is recommended in CKD trials because of its efficiency, control of the type 1 error rate, ethicality of all subjects receiving active treatment, and its appeal to patients. For any given type 1 error rate and power, the crossover design requires fewer patients than does a parallel design or delayed start design. The two period open label design is not recommended because it does not control the type 1 error rate in some scenarios. The delayed start design may be attractive to patients with conditions for which no effective treatment is available, because all patients in a given trial are guaranteed to receive the experimental therapy (either from the start or after the end of period 1). The crossover design also has this benefit. The delayed start design has considerably lower power than does either the parallel design or the crossover design.

Of the analysis strategies compared here, the mixed effects model analysis is expected to be most efficient if the assumed model is correct, and the subjects have various patterns of missing data. If all of the subjects have the same number of observations at each time point, then the mixed effects model is expected to be similarly efficient to the pooled analysis strategy. If there is a possibility of a large carryover effect, a two stage analysis may be more powerful than is pooled analysis. In the alternative scenario shown in Table 2, there was a moderate amount of carryover (25% of the treatment effect). In that scenario, two stage analysis and pooled analysis had approximately equivalent power. If the carryover effect were larger, the two stage test would have greater power.

One frequent concern in CKD trials is missing data. This issue can be mitigated by providing incentives for subjects to remain in a trial, even if they no longer wish to take the study drug. Regardless, a portion of subjects in any CKD trial will die, undergo kidney transplantation, or start dialysis. A concern with parallel trials is that the two groups may not be comparable after a large number of subjects is lost to follow up. However, in a crossover trial (in which each subject serves as his/her own control), the estimate of the treatment effect is not confounded by differences in covariate distribution between the two groups.

In future trials of CKD treatments, a two period design is recommended with an endpoint of the rate of change in eGFR. Our simulations and theoretical calculations generally agree with the empirical observations of Lathyris et al. [10]. Based on their review of meta-analyses (that included both crossover and parallel studies), Lathyris et al. concluded that crossover trials tend to agree with parallel arm trials. However, the group also found that parallel arm trials tended to make more conservative treatment effect estimates than did crossover trials. We also recommend a longer duration of total follow-up and a much smaller sample size in crossover

trials compared to those of parallel trials. Finally, a crossover trial may be ethically preferable to a parallel trial, because all of the subjects will receive the study drug. This feature may also facilitate subject recruitment.

Conflicts of interest

The author has no conflicts of interest to declare.

Acknowledgments

The authors thank Graeme A. O'May, PhD and J. Rick Turner, PhD, DSc of DRT Strategies, Inc. for editorial assistance.

ORCID

John Lawrence, https://orcid.org/0000-0002-9892-2753

References

1. Jha V, Garcia-Garcia G, Iseki K, et al. Chronic kidney disease: global dimension and perspectives. *Lancet* 2013;382:260–272.
2. Baigent C, Herrington WG, Coresh J, et al. Challenges in conducting clinical trials in nephrology: conclusions from a Kidney Disease-Improving Global Outcomes (KDIGO) Controversies Conference. *Kidney Int* 2017;92:297–305.
3. Inker LA, Heerspink HJL, Tighiouart H, et al. GFR slope as a surrogate end point for kidney disease progression in clinical trials: a meta-analysis of treatment effects of randomized controlled trials. *J Am Soc Nephrol* 2019;30:1735–1745.
4. Thompson A, Smith K, Lawrence J. Change in estimated GFR and albuminuria as end points in clinical trials: a viewpoint from the FDA. *Am J Kidney Dis* 2020;75:4–5.
5. Torres VE, Chapman AB, Devuyst O, et al. Tolvaptan in later-stage autosomal dominant polycystic kidney disease. *N Engl J Med* 2017;377:1930–1942.
6. D'Agostino RB Sr. The delayed-start study design. *N Engl J Med* 2009;361:1304–1306.
7. Lawrence J. Statistical review and evaluation: clinical studies (NDA 204-441) [Internet]. Silver Spring (MA): Center for Drug Evaluation and Research, 2017 [cited 2020 Sep 13]. Available from: https://www.accessdata.fda.gov/drugsatfda_docs/nda/2018/204441Orig1s000StatR.pdf.
8. Wang SJ, Hung HM. Use of two-stage test statistic in the two-period crossover trials. *Biometrics* 1997;53:1081–1091.
9. National Research Council. The prevention and treatment of missing data in clinical trials. Washington, DC: National Academies Press; 2010.
10. Lathyris DN, Trikalinos TA, Ioannidis JP. Evidence from crossover trials: empirical evaluation and comparison against parallel arm trials. *Int J Epidemiol* 2007;36:422–430.