Laura Cosgrove
Professor Ken Cheung
Bayesian Analysis and Adaptive Designs
December 13, 2019

Implications of Treatment-by-Patient Heterogeneity in Meta-Analysis of a Series of n-of-1 Trials

I. Introduction

Randomized controlled trials (RCTs) remain the most effective strategy to quell the statistical problems of confounding, identifiability, and bias that plague a would-be causal answer to a treatment benefit question. Common approaches taken by drug developers to minimize the risk of failure in the drug approval process include the graduation of competing potential agents through single-arm safety trials to preliminary efficacy trials with adaptive designs, spending the creativity of statistical approaches at earlier stages so as to choose the best agent with the most appropriately refined selection criteria that would show success in a "standard" RCT. However, criticisms of RCTs abound in those who advocate that the rigor, expense, and selection criteria of such a trial is at crossheads with practical realities in many clinical settings, which induces a lack of generalizability from RCT results to clinical practice (Gabler, Duan, Vohra, & Kravitz, 2011; Punja et al., 2016).

One increasingly popular approach is rooted in one of the earliest statistical designs in medicine and psychotherapy, the multiple crossover trial. Named "n-of-1 trial" in its more recent resurgence, this longitudinal design aims to evaluate the effect of one treatment contrasted against another for a single patient by alternating treatments over time. As collecting certain types of data becomes less expensive and less burdensome with the advent of better tools for remote monitoring and reporting such as mHealth and fitness trackers, a more-involved protocol including more measurements for a single person may be more feasible in some clinical settings than a pragmatic trial collecting a small amount of data on many people.

N-of-1 trials are best conducted in chronic, relatively stable conditions, where the treatment is quick in onset and, when removed, quick to cease its effect. These characteristics are related to important design considerations for an effective administration of a single n-of-1 trials, which include randomization of multiple blocks of a treatment sequence, to ensure trend or cyclical effects do not render the particular treatment inseparable from the trend and a washout period to eliminate carry-over effects from one treatment period (Punja et al., 2016). In an analogue of equipoise for clinical trials for a new treatment, there should also be doubt about a single "one-size-fits-all" approach to the condition[1] (Barr et al., 2015; Gabler et al., 2011). Otherwise, it would seem to be a comparative waste of resources to conduct multiple n-of-1 trials for each patient when a decision may be reached for all patients from a single RCT (or a review of the literature).

When appropriately and effectively run, N-of-1 trials have clear benefits for the participants under study. The approach is individualized; participants each have the opportunity to be on active treatment; participants will know their results quickly; and results will be, by definition of the question of identifying a treatment that works for each individual participant under study, relevant to the participant themselves, which is not always the case for an RCT

[1] That is, unlike in the setting of standard statistical analysis of two-group data in the hypothesis testing tradition such as Fisher's exact test, the assumption of uniform treatment effect for all hypothetical participants should be a poor assumption.

without sufficient subgroup analysis (Punja et al., 2016). However, the ethical question of whether n-of-1 trials are less beneficial than directing the same resources to an RCT for the broader patient population hinges on whether information can be aggregated from parallel or serial n-of-1 trials in order to answer the average treatment benefit questions that an RCT would answer. Though the evaluation of a series of n-of-1 trials can hinge on whether the n-of-1 trials successfully reach clinical decisions and an appropriate level of satisfaction for the participants and clinicians involved (Guyatt et al., 1990), or whether an n-of-1 trial approach is better for patients on average compared to standard clinical care (Barr et al., 2015), the same questions of treatment benefits or harms asked in an RCT or pragmatic trial can be asked of a series of n-of-1 trials through various meta-analysis approaches (Araujo, Julious, & Senn, 2016; Senn, 2019; D. R. Zucker et al., 1997; Deborah R. Zucker, Ruthazer, & Schmid, 2010). In fact, the Oxford Center for Evidence-Based Medicine names a systematic review of n-of-1 trials among the highest grade of evidence to evaluate treatment benefit, at the same grade as a systematic review of RCTs.

Three important statistical concepts in meta-analysis n-of-1 trials are individual treatment effects (ITE), average treatment effect (ATE), and heterogeneity of treatment effect (HTE). Above, ITE is discussed in terms of a binary individual treatment decision for a given patient in a single n-of-1 trial; this is the hypothesis-testing framework for an individual n-of-1 trial, whereas the estimation framework considers the ITE quantity as the “population” parameter of interest, where the “population” for a single individual might be considered the true individual effect from which any measured effect is a sample taking into account random variability of time and measurement. Similarly, in any general setting the ATE can be thought of as the either the binary treatment decision, from where we govern the operating characteristics of the analysis in terms of the probability of not rejecting the sharp null hypothesis of no ATE when there is, in fact, an ATE (Type II error), the probability of rejecting the sharp null hypothesis when there is no ATE (Type 1 error), and the probability of making a wrong treatment decision, a concept sometimes called Type III error. In the setting of estimating an ATE in a series of n-of-1 trials, the true population ATE is the parameter of interest, and the ATE we measure is a product of the random variability of time, of measurement error, of measurement across individuals due to individual differences in the outcome at baseline, and finally, the systematic variability due to the varying treatment effects depending on the individual. This final concept is the treatment-by-patient heterogeneity, or HTE, and unlike the sources of random variability which can be thought of as nuisance parameters only important to estimate as a means to estimate the ATE, the HTE is fundamental to both the motivation for an n-of-1 trial program and for possible further work in decomposing the source of treatment-by-patient heterogeneity due to the effects of patient characteristics, as in subgroup analysis.

Overall, this decomposition of variation in meta-analysis of n-of-1 trials is essential to the simulation study conducted by Araujo and colleagues in 2016, which served as the basis for my simulation study and interactive visualization (Araujo et al., 2016; Cosgrove, 2019). The thesis of the simulation project is the following: for population inference based on an analysis of many n-of-1 trials, the choice of method should best fit the features that motivate the choice of and design of n-of-1 trials, namely, the presence of HTE, and important design considerations such as the presence of cycle effects inducing measurement autocorrelation. In the work by Araujo and colleagues, the sources of variation modeled were cycle effects, occasion effects, heterogeneity of the population mean of the outcome in patients, and HTE.

II. Methods

The recommendations of Araujo et al. are summarized in the following: in the presence of treatment-by-patient heterogeneity and period heterogeneity, it is always valid and recommended to use a random effects model accounting for the above sources of variation, and when only summary measures are present it is also valid to use a random effects meta-analysis model for estimation. However, for simply testing the strict null hypothesis of no treatment effect, treatment interaction by patient can be ignored, and either matched-pair t tests or fixed effect meta-analysis can be used. I evaluate the Araujo et al. results in a new simulation study. I simulate the data using a mixed-effects model as outlined in Araujo et al., and my question of interest will be the operating characteristics of four analytic approaches in terms of rejecting the sharp null hypothesis of no average treatment effect, estimating the ATE, and estimating the HTE: the mixed-effects model approach mirroring that used to simulate the data; a matched-pair t-test; a random effects meta-analysis; and a fixed-effect meta-analysis.

III. Simulation Study

A. Setting

As a guide, I will construct the simulation study in the assumed setting of a meta-analysis of the PREEMPT study of chronic pain (Barr et. al, 2016), whose design but not result has been published, attending to important features of the design and data in order to construct my simulation dataset, assuming that the question of interest is the population average treatment effect of one treatment contrasted against one other. This research question is not the one that the PREEMPT study is trying to answer; the purpose of the PREEMPT study is not to estimate an average treatment effect, but rather to evaluate the strategy of conducting series of n-of-1 trials to reach an estimate of individual treatment effect versus clinical care as usual. To that end, even though the study is powered to recruit 122 patients in the n-of-1 arm, the patients A-B treatment contrast is not fixed ahead of time but rather can be any single agent or combination agent, including alternative therapies, contrasted against any other (for example, acetaminophen versus low-dose acetaminophen/hydrocodone, or low-dose acetaminophen/hydrocodone plus music therapy versus naproxen plus tramadol). Because of this design, it is highly unlikely that a hypothetical meta-analysis of the PREEMPT trial will be able to use 122 patients, so an optimistic hypothetical subset of n-of-1 trials with identical treatment contrast should be chosen[2]. One can think of the treatment contrast as not being strictly the particular agent, but for example, contrasting a medication-only approach with the medication-plus-alternative therapy approach, in order to include more patients in the analysis. However, the more general the characteristics of the contrast, the more likely a large treatment-by-patient heterogeneity will need to be modeled. To that end, I chose the sample size following the first simulation of a clinically-relevant average treatment effect with no heterogeneity as the sample size with 90% power to reject the sharp null hypothesis under the matched pair t test.

Other parameters chosen is the population mean of the outcome, which are NIH PROMIS scores normalized to have a mean of 50 points in a clinically-relevant population; the population standard deviation of the outcome, the source of variation being by-patient heterogeneity, which is normalized to have a standard deviation of 10 points; cycle effects and occasion effects, which

[2] A more complete and formal analysis might include information from every n-of-1 trial that contains one of the contrasts in the forthcoming PREEMPT data, analogous to using Hierarchical Bayesian analysis to borrow information about observational studies, but this approach is for future directions.

were chosen to have a standard deviation of 10 points; and, finally, average treatment effect of the lower bound of clinical relevance of 4 points, with a very large HTE of 16 points when THE was included.

B. Approach

The model to simulate the data is as follows:

$$Y_{irs} = \lambda_i + \beta_{ir} + \epsilon_{irs} + Z_{irs}\tau_i$$

where Y_{irs} is the measured outcome for occasion *s*, $s \in (1, 2)$ of cycle *r*, $r \in (1,2,....k)$ for patient *i*, $i \in (1,2, .. n)$, and:

- $\lambda_i \sim N(\Lambda,\phi^2)$: Patient-specific mean of the outcome, drawn from a normal distribution with population mean of Λ
- $\beta_{ir} \sim N(0,\gamma^2)$: Within patient, within cycle (block) random noise
- $\epsilon_{irs} \sim N(0,\sigma^2)$: Within patient, within cycle (block), within occasion (independent of treatment) random noise
- $\tau_i \sim N(T,\psi^2)$: Where Z_{irs} takes values ½ when the occasion *s* has treatment A and – ½ when occasion *s* has treatment B, the patient-specific for treatment effect, drawn from a normal population mean treatment effect of T. Therefore, an estimate for HTE (heterogeneity of treatment effect) is ψ.

Mixed Effects Model

The mixed effect model is the same as the underlying model that generated the data, and is fit on all data. It estimates a cycle effect nested within patients, and a random patient effect with a slope term of treatment effect. We recover our parameter values in the mixed effects model as follows:

- We set ϕ in $\lambda_i \sim N(\Lambda,\phi^2)$, as 10 points. In the model, the corresponding estimate in the variance of the outcome is under the estimated standard deviation of the random effect of the individual id.
- We set Λ in $\lambda_i \sim N(\Lambda,\phi^2)$ as 50 points. In the model, the corresponding estimate in the mean of the outcome is under the fixed effects intercept term.
- We set γ in $\beta_{ir} \sim N(0,\gamma^2)$ as 10 points for most of our simulations. In the model, the corresponding estimate in the mean of the outcome is under the estimated standard deviation of the nested random effect, id:cycle.
- We set σ in $\epsilon_{irs} \sim N(0,\sigma^2)$ as 10 points for most of our simulations: in the model, it is the estimated standard deviation of the residual random effect.
- We set T in $\tau_i \sim N(T,\psi^2)$, as 4 points for most scenarios. In the model, it is estimated under the fixed effects slope term for Z.
- We set ψ in $\tau_i \sim N(T,\psi^2)$, as 16 points for most scenarios. In the model, it is estimated under the estimated standard deviation of the random effects slope term of Z by treatment.

Matched Pairs T Test

The matched pair t-test is performed on each pair of within-cycle treatment outcomes for each patient, and the *nk* matched outcomes are treated as independent. Note that here, and for the summary measures approach for meta-analysis, if there were multiple measurements per cycle the mean of each treatment within cycle was taken as the pair difference.

Random Effects Meta-Analysis

Unlike the previous two approaches, in the meta-analysis realm, it is assumed that one does not have access to the n-of-1 studies' raw data, but rather summary measures for each study. For both random and fixed effects meta-analysis, Araujo et al. showed that that the summary measures of the mean and variance of within cycle differences for n-of-1 studies, or in other words, the ITE and its variance, is a basic estimator approach that estimates the average treatment effect and variance of the treatment effect (treatment heterogeneity) correctly. When random effects meta-analysis is performed on these summary measures, the *rma.uni* function from the *metafor* package is used. Both "total treatment heterogeneity" and the treatment effect is estimated. I take the total treatment heterogeneity estimate as that of heterogeneity of treatment effect.

Fixed Effects Meta-Analysis

When fixed effects meta-analysis is performed on these summary measures, the *rma.uni* function from the *metafor* package is used with method "*FE*". The fixed effect approach weights the studies' estimated ITE inversely to their estimate of its variance in a simple weighted average.

Operating characteristics

To evaluate the performance of the various analytic approaches, 1000 independent replicates of series of n-of-1 trials were simulated under the same population parameters. The below table describes the simulations. Simulation 1 was constructed in order to detect a clinically relevant, uniform effect, and to choose the sample size for further simulations based on a 90% power threshold for the matched pair t-test. Simulation 2 was constructed to detect a clinically relevant, heterogeneous effect. Simulation 3 was constructed in order to increase the number of patients in the Simulation 2 scenario to the maximum sample size in the PREEMPT trial, in hopes of reaching the same power to reject the sharp null as Simulation 1. Simulation 4 was constructed to contrast these power gains with the strategy of increasing the number of cycles in Simulation 2 to a practical maximum based on the PREEMPT design description. Simulation 5 was constructed to evaluate the Type I error under the null scenario, with no HTE, and simulation 6 was constructed to evaluate the Type I error under the null scenario with large HTE.

	1	*2*	*3*	*4*	*5*	*6*
ATE, T	4	4	4	4	0	0
HTE, ψ	0	16	16	16	0	16
N	34	34	122	34	34	34
K	4	4	4	8	4	4
S	2	2	2	2	2	2
measurements	1	1	1	14	1	1
γ	10	10	10	10	10	10
σ	10	10	10	10	10	10
Λ	50	50	50	50	50	50
ϕ	10	10	10	10	10	10

IV. Results

Simulation 1

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*
Mixed Effects Model	0.90	0.00	0.96	0.95	2.10
Matched Pair T Test	0.91	0.00	0.95	0.95	NA
Random Effects Meta-Analysis	0.67	0.00	0.94	1.35	0.90
Fixed Effects Meta-Analysis	0.67	0.01	0.81	1.83	NA

Simulation 2

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*
Mixed Effects Model	0.29	0.00	0.95	2.38	1.90
Matched Pair T Test	0.57	0.00	0.77	2.38	NA
Random Effects Meta-Analysis	0.26	0.00	0.94	2.52	3.16
Fixed Effects Meta-Analysis	0.67	0.10	0.43	4.51	NA

Simulation 3

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*

Mixed Effects Model	0.71	0.00	0.94	1.21	1.02
Matched Pair T Test	0.91	0.00	0.80	1.21	NA
Random Effects Meta-Analysis	0.68	0.00	0.95	1.28	2.45
Fixed Effects Meta-Analysis	0.85	0.08	0.38	3.20	NA

Simulation 4

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*
Mixed Effects Model	0.32	0.00	0.94	2.18	1.59
Matched Pair T Test	0.81	0.02	0.52	2.18	NA
Random Effects Meta-Analysis	0.32	0.00	0.94	2.18	1.70
Fixed Effects Meta-Analysis	0.88	0.07	0.28	2.74	NA

Simulation 5

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*
Mixed Effects Model	0.05	0.00	0.95	1.00	2.09

Matched Pair T Test	0.06	0.00	0.94	1.00	NA
Random Effects Meta-Analysis	0.06	0.00	0.94	1.50	0.85
Fixed Effects Meta-Analysis	0.20	0.00	0.80	1.88	NA

Simulation 6

Method	*Proportion where sharp null was rejected*	*Proportion where the wrong treatment was recommended*	*Proportion where the CI contained true treatment effect*	*MSE, ATE*	*MSE, HTE*
Mixed Effects Model	0.07[3]	0.00	0.93	2.46	1.89
Matched Pair T Test	0.25	0.00	0.75	2.46	NA
Random Effects Meta-Analysis	0.08	0.00	0.92	2.60	3.13
Fixed Effects Meta-Analysis	0.56	0.00	0.44	4.66	NA

V. Discussion

Overall, the recommendations of Araujo and colleagues were validated in this simulation study. Specifically, in the presence of treatment by patient heterogeneity, models that include a random effect always estimate the ATE more accurately. Random effect models are less powerful for testing the sharp null hypothesis, but random effect models better preserve Type 1 error for the sharp null hypothesis when there is treatment by patient heterogeneity, given that the matched pair t-test underestimates the true variances of the ATE estimates. Fixed effects meta-analysis seems to have poor operating characteristics for both estimation and hypothesis testing. And in most cases, except when the HTE is 0, the mixed effects model run on the underlying data better estimates the HTE.

[3] This should, theoretically, be fixed at 0.05 even when heterogeneity is present given that it is the true model. A possibility is that the number of random effects is too large, leading to singularity problems in estimation, or the number of simulations remained too low.

Increasing the number of measurements and cycles did marginally improve power of the random effects approaches in the presence of HTE, but while it improved power of the matched pair t-test it worsened estimation, likely because the number of paired differences inflated the certainty of the test while the variation due to HTE remained a source of bias due to a small patient sample. However, increasing the number of patients improved power of the random effects models by about 40%. Comparing the overall power in the presence of HTE, the sharp null hypothesis was rejected in 90% of series both when the series contained 122 patients and an HTE of 16 compared to 34 with an HTE of 0 and the matched pair t test was used. However, estimation performance was worse. The mixed effects model, in contrast, had worse power to reject the sharp null hypothesis when HTE was present but estimation remained effective at 95%.

The feasibility of conducting such a meta-analysis on the PREEMPT trial is unclear. For the analytic approaches that are valid for estimation, it is unlikely they would be powered to conclusively detect an average treatment effect if the parameter values set are taken as truth. However, a standard deviation of 16 for HTE is quite large. Including covariates in an analysis may account for the heterogeneity such that the residual HTE unexplained by the covariates is smaller, resulting in a more powerful analysis. Further work in sample size considerations for n-of-1 trials in a similar analytic setting is described in a 2019 article by Stephen Senn (Senn, 2019).

VI. Future Directions

While Araujo et al. do not discuss Bayesian hierarchical models, Empirical Bayes estimates are discussed in a further important conclusion advancing the use of a mixed-model approach to estimate the variance of the treatment effect. That is, if one were to use the results of the n-of-1 meta-analysis to treat future patients, one can use the estimated treatment effect mean and the treatment effect variance to improve the interim treatment effect estimates for a patient using the Empirical Bayes method for example after one cycle. It seems to me that a fully Bayesian treatment may also be used for future patients; the essential ingredient is that the heterogeneity of the treatment effect is estimated with the meta-analysis.

As a contrast, Zucker et al. (2010) give an application of n-of-1 meta-analysis comparing mixed effects models on the underlying measurements, fixed effect and random effect meta-analysis on the summary patient measures, and Bayesian hierarchical models that I find perfectly thorough, but ultimately somewhat disappointing due to the fact that the for data in question seemed to lack HTE, an essential motivation for an n-of-1 trial program. The Zucker paper also showed a characteristic difficulty of the application of hierarchical Bayesian models, or Bayesian analysis in general: the task of specifying the priors for the within-patient variance of measurements and between-patient variances of the effect of the control treatment and the difference between the control and standard treatment. With no good reason to choose an informative prior, one might as well forgo the Bayesian analysis given that the models under a noninformative prior are similar (Zucker et al. 2010). However, for future n-of-1 trials where the extent of heterogeneity is relatively well-established, the same Bayesian priors may be used for meta-analysis and individual n-of-1 trial treatment decision, a theoretical consistency that is appealing.

Yet there is an important discussion not in Araujo et al. that Zucker et al. account for (though ultimately, it was not important in their application): the explanatory relationship of other covariates, either exercising their effect on the response through a population (fixed) effect or through a random effect (a covariate-by-patient interaction). The reason I think this is an essential discussion that is in line with what I take as the philosophy of Araujo et al., that n-of-1 analysis should account for the purposes of an n-of-1 design, is due to the fact that a natural extension of the assumption of individual treatment heterogeneity is the assumption that some, measurable covariates may explain individual treatment heterogeneity. Given that the hope of any treatment effect estimation would be to affect future care decisions, and in the hope of personalizing medicine, estimation of heterogeneity by some explanatory covariates that would provide more precise prior estimates of treatment effect for future patients belonging to strata defined by those covariates seem an important strength of the more clinical-care-oriented n-of-1 study contrasted with an RCT.

VII. General Reflection

I am very glad I took this class. I feel that I am coming away from the course with an understanding of how to evaluate new trial designs in terms of their operating characteristics, and an appreciation for the Bayesian philosophy toward estimation that seems essential to conceptualize certain adaptive designs and, to a large extent, many types of simulation studies. Will I go on to remember the particular characteristics of every trial design we discussed, given that I'm not expecting to work in the clinical trial space immediately after graduation? Probably not, but in every health industry or enterprise, there is experimentation, and I will hopefully be better prepared to think creatively about which methods of experimentation makes sense for a given experimental goal. Thank you, Dr. Cheung and Ziwei!

References

Araujo, A., Julious, S., & Senn, S. (2016). Understanding Variation in Sets of N-of-1 Trials. *PLOS ONE*, *11*(12), e0167167. https://doi.org/10.1371/journal.pone.0167167

Barr, C., Marois, M., Sim, I., Schmid, C. H., Wilsey, B., Ward, D., … Kravitz, R. L. (2015). The PREEMPT study - evaluating smartphone-assisted n-of-1 trials in patients with chronic pain: Study protocol for a randomized controlled trial. *Trials*, *16*(1), 67. https://doi.org/10.1186/s13063-015-0590-8

Cosgrove, L. (2019, December 1). n-of-1 Design Exploration. Retrieved from https://lauracosgrove.shinyapps.io/nof1/

Gabler, N., Duan, N., Vohra, S., & Kravitz, R. (2011). N-of-1 Trials in the Medical Literature: A Systematic Review. *Medical Care*, *49*(8), 761–768. https://doi.org/10.1097/MLR.0b013e318215d90d

Guyatt, G. H., Keller, J. L., Jaeschke, R., Rosenbloom, D., Adachi, J. D., & Newhouse, M. T. (1990). The n-of-1 randomized controlled trial: Clinical usefulness. Our three-year experience. *Annals of Internal Medicine*, *112*(4), 293–299. https://doi.org/10.7326/0003-4819-112-4-293

Punja, S., Bukutu, C., Shamseer, L., Sampson, M., Hartling, L., Urichuk, L., & Vohra, S. (2016). N-of-1 trials are a tapestry of heterogeneity. *Journal of Clinical Epidemiology*, *76*, 47–56. https://doi.org/10.1016/j.jclinepi.2016.03.023

Senn, S. (2019). Sample size considerations for n-of-1 trials. *Statistical Methods in Medical Research*, *28*(2), 372–383. https://doi.org/10.1177/0962280217726801

Zucker, D. R., Schmid, C. H., McIntosh, M. W., D'Agostino, R. B., Selker, H. P., & Lau, J. (1997). Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. *Journal of Clinical Epidemiology*, *50*(4), 401–410. https://doi.org/10.1016/S0895-4356(96)00429-5

Zucker, Deborah R., Ruthazer, R., & Schmid, C. H. (2010). Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: Methodologic considerations. *Journal of Clinical Epidemiology*, *63*(12), 1312–1323. https://doi.org/10.1016/j.jclinepi.2010.04.020