Individual \((N\text{-of-1})\) trials can be combined to give population comparative treatment effect estimates: methodologic considerations

Deborah R. Zucker\(*\), Robin Ruthazer, Christopher H. Schmid

Institute for Clinical Research and Health Policy Studies, Tufts Medical Center, Tufts University, Boston, MA, USA

Accepted 4 April 2010

Abstract

Objective: To compare different statistical models for combining \(N\text{-of-1}\) trials to estimate a population treatment effect.

Study Design and Setting: Data from a published series of \(N\text{-of-1}\) trials comparing amitriptyline (AMT) therapy and combination treatment (AMT + fluoxetine [FL]) were analyzed to compare summary and individual participant data meta-analysis; repeated-measure models; Bayesian hierarchical models; and single-period, single-pair, and averaged outcome crossover models.

Results: The best-fitting model included a random intercept (response on AMT) and fixed treatment effect (added FL). Results supported a common, uncorrelated within-patient covariance structure that is equal between treatments and across patients. Assuming unequal within-patient variances, a random-effect model was favored. Bayesian hierarchical models improved precision and were highly sensitive to within-patient variance priors.

Conclusion: Optimal models for combining \(N\text{-of-1}\) trials need to consider goals, data sources, and relative within- and between-patient variances. Without sufficient patients, between-patient variation will be hard to explain with covariates. \(N\text{-of-1}\) data with few observations per patients may not support models with heterogeneous within-patient variation. With common variances, models appear robust. Bayesian models may improve parameter estimation but are sensitive to prior assumptions about variance components. With limited resources, improving within-patient precision must be balanced by increased participants to explain population variation. © 2010 Elsevier Inc. All rights reserved.

Keywords: \(N\text{-of-1}\) trials; Methodology; Comparisons; Population estimate; Meta-analysis; Comparative effectiveness

1. Introduction

\(N\text{-of-1}\) trials are often multi-crossover, randomized, blinded single-patient trials, in which an individual generally tries two interventions (two treatments or a treatment and a placebo), each multiple times, to determine which is more effective [1]. By providing more rigorous comparative effectiveness assessments for an individual, these trials may offer an approach for providing evidence-based, personalized medicine. Classically, these single-patient \((N=1)\) trials were designed to aid personal clinical decision making and to focus solely on the individual’s response to treatments. Thus, the frequency of \(N\text{-of-1}\) trial use in current practice is not known, because individual care—focused trials comparing approved treatments and undertaken solely for personal treatment management are not typically published. Most published reports of \(N\text{-of-1}\) trials summarize and compare results across trials from a prospectively designed series. We have proposed combining series of similarly designed \(N\text{-of-1}\) trials using a Bayesian hierarchical model for comparative effectiveness research [2]. In addition to incorporating external information, this methodology gives both a comparative treatment effect estimate for the population and potentially improved effectiveness estimates for individuals by borrowing strength from the trials of others to learn about between- and within-patient variations. The \(N\text{-of-1}\) trial’s presumption of individual uniqueness and its goal of personal prediction would seem to contrast with the usual clinical trial objective of identifying a common treatment effect to then apply to each individual in the population. Because combining a series of \(N\text{-of-1}\) trials can provide an estimate of overall effect, it is important to understand when one might use this approach and how this approach and its resulting estimate might agree or differ from more standard clinical trial designs.

\(N\text{-of-1}\) trials have been used to compare interventions for a range of chronic conditions (e.g., arthritis [3], fibromyalgia [4,5], chronic airflow limitation [6], gastric reflux [7], and
What is new?

Key findings
In combining N-of-1 trials to estimate a population treatment effect:

- With few observations per patient and little information about within-patient variation, combined N-of-1 trials data may not support models that include complex variance structures.
- Prior information with Bayesian models can be useful for increasing the precision of estimates but are very sensitive to prior assumptions about variance components.
- Models with fixed treatment effects and common variances are robust and lead to conclusions that are similar to, though more precise than, single-period or single-crossover study designs.

What this adds to what was known:
- N-of-1 studies can be combined to estimate population effects and, depending on the analytic models used, might offer different estimates and insights compared with more standard clinical trials.

What is the implication? What should change now?
- Combining N-of-1 trials provides an approach for estimating population treatment effects; although with limited resources, improved within-patient precision from repeated measures must be balanced by increased participants to explain population variation, to give representative population effect estimates and to address patient care and research goals.

Completed trials, what is the best way to combine their data? If one is prospectively collecting data, do additional measures per patient increase the precision of the population estimate? Do repeated measurements on individuals allow for models that are better than those based on one or two measurements per patient? Although improved estimation of each patient’s individual response is a prime advantage of N-of-1 trials and having these embedded within population trials might provide implementation benefits [12], we focus on the analytic considerations and models for estimating the population effect in this article.

Each N-of-1 trial provides information not only about within-patient treatment comparisons but also about the variation of responses within each patient to each treatment. Therefore, we wanted to assess whether and how this within-patient variation might best be incorporated into a population estimate that presumably depends more on between-patient than within-patient variation. To combine N-of-1 trials, one might first follow the summary meta-analysis paradigm and combine results within each (patient’s) trial to estimate individual patient effects and then average across (patients’) trials to obtain an overall effect [13,14]. This type of analysis can use available data either by patient period or aggregated to the patient level (e.g., if one has a two-period trial, one would use the simple within-subject difference between the two treatments’ outcomes). Alternatively, with patient-period data, as with individual patient data meta-analysis, one could use mixed models to combine within- and across-patient periods simultaneously [15–18]. For prospectively designed trials with treatment order randomized across trials, the use of only a subset of the periods would mirror more standard population designs. For example, analyzing randomized first-period treatment outcomes corresponds to a randomized parallel group trial; using outcomes from the pair-randomized treatments in the first two periods corresponds to an AB/BA crossover design. However, not all trial series are prospectively designed and randomized together for combined N-of-1 analysis. Comparing the subset analyses with models that use all periods can help evaluate the information added by the multiple crossovers (obtained at higher cost to participants and to the health care/research system).

To inform our comparisons of the various models for optimally combining N-of-1 trials, we use data from our study of N-of-1 trials comparing amitriptyline (AMT) and the combination of AMT and fluoxetine (FL) for treating fibromyalgia syndrome (FMS) [5]. By reanalyzing the data from this series of similarly designed N-of-1 trials, we compare and contrast meta-analysis, repeated-measure models, Bayesian hierarchical models, and single-period and single-pair crossover models for combining N-of-1 trials to obtain a population estimate of comparative treatment effectiveness. We consider key sources of variation (within patient, between patient, and random) and the assumptions of these different models, and use our data to aid in choosing between the various analytic options.
2. Methods

2.1. Data collection and selection

The data come from a series of 58 N-of-1 trials that compared two therapies for FMS and were carried out at seven community-based rheumatology practices and at one FMS referral center [5]. Each N-of-1 trial had six treatment periods divided into three sets of paired treatments—one period on combination therapy (AMT + FL) and one on the “control” medication (AMT). For treatment allocation, the first treatment pair was block randomized (for every 20 trial kits, 10 started with AMT and 10 with AMT + FL). The other two pairs were random in their start mediciations. The Fibromyalgia Impact Questionnaire (FIQ) score [19] was the outcome measured once at baseline (absent FMS medications) and once at the end of each of the six 6-week treatment periods. The FIQ is a patient-assessed quality-of-life assessment tool, scored continuously between 0 (best score) and 100 (worst score). Analyses reported here used data from the 46 patients who completed at least two treatment periods (one on each treatment), of whom 34 completed all six treatment periods [5].

2.2. Analytic approaches for combining N-of-1 trials

We investigated three analytic methods for combining all data available from the N-of-1 trials: (1) summary data meta-analysis, (2) linear mixed models using maximum likelihood (individual participant data [IPD] meta-analysis); and (3) Bayesian hierarchical models with and without informative prior information. We compared these answers with analyses applied to portions of the data that mirrored more standard clinical trial designs: (1) a t-test applied to only the first-period measurement from each trial (analogous to a randomized parallel-group trial); and (2) a paired t-test of data from only the first two periods (one on each treatment) of each N-of-1 trial (analogous to a typical AB/BA single-crossover design). We also used a paired t-test to estimate the mean difference of each patient’s averaged outcome on each treatment across the N-of-1 trials. To simplify model comparisons, we focus here on analyses using data from the 34 individuals who completed their N-of-1 trials (all six treatment periods) and, thus, have balanced trial designs. Later, we discuss some results obtained using data from all the individuals \((N = 46)\) with at least two completed periods (one on each treatment) in beginning to explore analyses of unbalanced N-of-1 trial designs.

2.2.1. Summary data meta-analyses: considering each N-of-1 study as a separate trial

A typical estimate of the pooled treatment effect combines summary data from each trial to form a weighted average \(\hat{\mu} = \frac{\sum_{i=1}^{N} w_i y_i}{\sum_{i=1}^{N} w_i}\) of the individual study estimates, \(y_i\), with weights, \(w_i\). In a fixed-effect model, one assumes that each study has the same true mean effect, \(\mu\), whereas in a random-effect model, each study can have a different effect, \(\mu_i\), but these effects are assumed to follow a common distribution, often assumed normal \(N(\mu, \tau^2)\). In a fixed-effect model, the studies’ weights \(w_i = 1/\sigma_i^2\) are often supplied by the inverse of their variances \(\sigma_i^2\); in a random-effect model, the weights \(w_i = 1/(\tau_i^2 + \sigma_i^2)\) also include a between-study variance component \(\tau^2\). Because this between-study variance component is the same for all the studies being aggregated, the resulting weights are more equal than in the fixed-effect model and are less sensitive to large differences among their within-study variances.

Using this summary data meta-analysis approach, a set of N-of-1 trials could be combined using the studies’ observed mean effects \(y_i\) and variances \(\sigma_i^2\). However, unlike a typical summary data meta-analyses in which large population trials supply good estimates of within-trial variances, N-of-1 studies typically have only a handful of observations, and hence, their within-trial variances are not well estimated.

Many years ago, Cochran [20] suggested that, when combining results of small population studies, it may be better to assume a common within-study variance formed by pooling across trials. Within-study inverse variance weights are then proportional to the studies’ sample sizes. When the trials’ designs are balanced and have the same number of unit measurements, their weights are equal. We fit fixed- and random effects—combined N-of-1 summary meta-analysis models with both separate and common variances to the combined N-of-1 data. Within-study variance was estimated using a standard sample variance formula. The between-study variance component was estimated using the method of DerSimonian and Laird [13], and the models were fit using S-Plus 6.2 (Insightful Corp, Seattle, WA).

2.2.2. Linear mixed models: combining IPD across trials

Increasingly, meta-analyses use mixed models to combine IPD across trials [14–17] while accounting for correlations in the data that derive from the clustering of participants within the trials. When these clusters (studies) are treated as a random sample from a population of potential clusters, the model contains additional variance components to represent the between-study variation. In typical IPD meta-analyses, the participants are the units providing the correlated measurements within the study clusters. In contrast, in a series of N-of-1 trials, the correlated data are the multiple comparative measurements taken on individuals, and the model then has within-patient variance components, as in longitudinal studies [18,21]. If these patients are then nested within a clinical practice or a specific trial series, the correlated data structure may be extended further to reflect this.

To use a linear mixed model to aggregate trials’ data requires clearly defined assumptions about the model’s multiple components and their relationships. In combining our N-of-1 trials, these include (1) determining the model parameters; (2) characterizing parameters’ effects on the
outcome (fixed- vs. random-effect assumptions); and (3) defining assumptions regarding the several variance components.

2.2.2.1. Model parameters. We let \( Y_i \) be the observed FIQ score for the \( i \)th patient in the \( j \)th period (\( j = 1, 2, \ldots, J \)) and let \( Y_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iJ}) \) denote the collection of FIQ scores for the \( i \)th patient. We assume that \( Y_i \) follows a multivariate normal distribution with mean \( \mu_i \) and a \( J \times J \) covariance matrix \( \Sigma_i \), expressed as \( Y_i \sim N(\mu_i, \Sigma_i) \). In the balanced repeated crossover design in our study, half of the measurements were taken under AMT + FL (treatment) and half under AMT (control). The mean treatment effect vector \( \mu_i \) then consists of separate values for the two treatments. Using an indicator variable \( X_{ij} = 1 \) for treatment with AMT + FL and \( X_{ij} = 0 \) for AMT, we write each individual period mean as \( \mu_{ij} = \alpha_i + \beta_i X_{ij} \), where \( \alpha_i \) is the control mean (mean FIQ score on AMT alone) and \( \beta_i \) is the difference between the two treatments’ means. In addition to the treatment effect, other covariates (such as trial period, practice setting, and participant characteristics) might be included in the model. To incorporate these variables, we can generalize the mean to \( \mu_i = X_i \theta_i \) (a function of a set of regression covariates \( X_i \) with a vector of regression parameters, \( \theta_i \)). In the single-treatment model, \( \theta_i = (\alpha_i, \beta_i) \).

2.2.2.2. Mean parameter specifications (fixed vs. random effects). In this model, when \( \theta_i \) is the same for each patient (\( \theta_i = \theta \)), \( \theta \) is a fixed effect. Alternatively, if the regression varies across patients, \( \theta_i \) would be a random effect, as for example, if each patient’s response were to depend on many patient-specific characteristics that affect the outcome in both treatment groups. It is expressed in a general form as \( \theta_i = \theta + \delta_i \), where \( \theta \) is the fixed effect and \( \delta_i \sim N(0, D) \) is a random effect with covariance matrix \( D \). The normal distribution for \( \delta_i \) specifies the form of this random variation across individuals. Effects with only a fixed component have the corresponding elements in \( D \) set to zero.

2.2.2.3. Multilevel model structures. Another way of representing this linear mixed model is to use a hierarchical or multilevel form:

\[
Y_i = X_i \theta_i + e_i \quad \text{with} \quad e_i \sim N(0, \Sigma);
\]

\( \theta_i \sim N(\theta, D) \).

As an example, if the outcome depends only on treatment, we can write this model as follows:

\[
Y_{ij} = \alpha_i + \beta_i X_{ij} + e_{ij} \quad \text{with} \quad e_{ij} \sim N(0, \sigma^2_i)
\]

\( \alpha_i \sim N(\alpha, \tau^2_\alpha) \)

\( \beta_i \sim N(\beta, \tau^2_\beta) \)

In specifying how each of the different subjects’ models relates to each other, this structure reduces the number of parameters that need to be fit for estimation while still allowing the mean (and possibly variance) to vary across individuals. Estimating the random intercepts \( \alpha_i \) and the random slopes \( \beta_i \), requires only two mean parameters (\( \alpha \) and \( \beta \)) and three variance parameters (\( \sigma^2, \tau^2_\alpha, \) and \( \tau^2_\beta \)). An additional correlation between \( \alpha_i \) and \( \beta_i \) may also be added. The random-effect summary meta-analysis model is a form of this model with a single summary statistic (e.g., the difference between the mean of the treatment and the mean of the control groups) as a univariate outcome. The treatment effect \( \beta_i \) has mean \( \beta \) (the overall treatment effect) and between-patient variance \( \tau^2_\beta \) and a within-patient variance \( \sigma^2_i \), which vary across patients. The fixed-effect model follows by setting \( \tau^2_\beta = 0 \).

2.2.2.4. Variance structures. Covariance matrices may be structured within patients and may vary across patients. When few observations are available from each patient, the same within-patient covariance matrix, \( \Sigma_i \), for all patients is often assumed, that is, \( Y_i \sim N(\mu_i, \Sigma) \). We investigated several forms for \( \Sigma \), including (1) unstructured (general \( \Sigma \)); (2) first-order autoregressive, in which each measurement has the same variance and measurements that are \( k \) visits apart have correlation \( \rho^k \); (3) uncorrelated errors with common variance \( \sigma^2 \) at each period; and (4) uncorrelated errors with separate variances for measurements on each treatment. The unstructured form, although the most general, requires fitting \( J \) (in our completed trial, \( J = 6 \)) variance and \( J(J + 1)/2 \) (i.e., 21) covariance terms and, hence, may be overparameterized. The first-order autoregressive form reflects the longitudinal nature of the short time series of measurements taken on each individual and uses two parameters, \( \sigma^2 \) and \( \rho \). The common variance form (across patients) with uncorrelated and identically distributed errors is the simplest possible structure, requiring only one parameter, \( \sigma^2 \). A variant of this common form allowing separate variances by treatment group can account for the possibility that response variability differs by treatment.

Although most mixed models assume a common within-patient covariance matrix across patients, if sufficient within-patient data are available, it may be possible to remove this restriction. The simplest model assumes an uncorrelated common-variance structure for each patient with different variances across patients, that is, \( \Sigma_i = \sigma^2 I \). The number of variances to be estimated may be reduced by grouping them according to some characteristic of the patients. For example, if we thought that FIQ scores for patients treated in the community might vary differently from those of patients treated at the referral center, we could set \( \Sigma_i = \Sigma_A \) for community-based patients and \( \Sigma_i = \Sigma_B \) for patients treated at the referral center. A simple form reflecting no within-person correlation and a common variance across measurements within a group of patients has \( \Sigma_A = \sigma^2_A I \) and \( \Sigma_B = \sigma^2_B I \).
In model development, the order of testing assumptions can impact model structure. For our models, we first considered whether factors besides the patient and the treatment significantly impacted treatment effect. These variables (in our example, time and site) were selected, because they were integral to our prospective study design [5]. We fit the linear mixed models using the lme function in S-Plus 6.1. Model comparisons used the Bayesian information criterion (BIC [22]) that penalizes the likelihood for the addition of parameters. BIC has been shown to maximize the posterior probability that one model out of a considered set is correct under the assumption that no models are preferable a priori. Models with lower BIC are preferred, because they provide the best fit with the fewest number of parameters.

2.2.3. Bayesian hierarchical models

Bayesian methods enable the incorporation of prior information regarding the treatment effect and inference about the resulting posterior probability estimates. The Bayesian version of the mixed model uses the same hierarchical structure while adding prior distributions on the model parameters. For a model with treatment as the sole covariate, this involves the addition of prior distributions for \( \alpha, \beta, \sigma^2, \tau^2_a, \) and \( \tau^2_b \). We derived [5] the prior distributions from a published crossover trial [23] that used the same medications and dosages. That trial’s results suggested a mean of 8 points and a standard deviation of 3.9 for \( \beta \), the mean difference between FIQ scores on AMT + FL and AMT, and a mean of 50 points with a standard deviation of 4.6 for \( \alpha \), the mean response on AMT (“control” treatment).

We had limited information about the variance parameters \( \sigma^2 \) (between patients), and \( \sigma^2_i \) (within patient). Inverse gamma (IG) distributions are convenient computational choices for distributions of variances, and we chose a fairly noninformative IG (1,8) prior distribution for all of these variance parameters. In addition to assuming random treatment effects, we also tested models assuming constant treatment effects across patients (\( \beta = \beta \)) by setting \( \tau^2_b = 0 \) and a common within-patient variance model (\( \sigma^2_i = \sigma^2 \)) with an IG (1,8) prior on \( \sigma^2 \).

To extend the Bayesian models for comparison with the non-Bayesian mixed model analyses and to assess the impact of the prior distributions, we also examined these models using noninformative prior distributions for all parameters. We chose normal distributions centered at 0 with very large variances of \( 10^6 \) for the means, \( \alpha \) and \( \beta \), and chose uniform distributions over the range \( 0 \rightarrow 10^4 \) for square roots of the variance parameters [24]. Bayesian models were fit using Markov chain Monte Carlo with the WinBUGS 1.4.3 software (available at: http://www.mrc-bsu.cam.ac.uk/bugs/winbugs/contents.shtml; Medical Research Council and Imperial College of Science, Technology and Medicine).

2.2.4. Standard trial design analyses

We compared these more complex models with standard population research design analyses. We mimicked a parallel-arm randomized controlled trials by comparing outcomes from only the first treatment period from each N-of-1 trial (AMT or AMT + FL) using a t-test. We analyzed outcomes from the first pair of periods (one period on each treatment) from each N-of-1 trial using a paired t-test as in an AB/BA crossover design [25]. We then also used a paired t-test to compare the difference in each individual’s averaged outcome from all completed periods on each treatment from across all the N-of-1 trials, which we here termed the averaged crossover design.

2.2.5. Additional considerations: more measures vs. more patients

Finally, we compared the impacts of adding treatment periods and patients into a series of N-of-1 trials. Using a simple random-effect model [14], we calculated the approximate precision (inverse of the variance) of the mean effect with \( M \) patients and \( N \) paired treatment periods per patient compared with a classic two-period (AB/BA) crossover design (\( M = 46 \) and \( N = 1 \)) under several combinations of between- and within-patient variances. Precision was calculated as the sum of the weights on individual patients where the weight on an individual was the sum of the within-patient and the between-patient variances (\( \sigma^2 + \tau^2 \), respectively). Patients are assumed independent; hence, the precision is calculated as \( M(\tau^2 + \sigma^2/N) \).

3. Results

We used the data from 46 N-of-1 trials with at least two completed treatment periods (one on each therapy) from our combined N-of-1 implementation study [5]. Among the 34 completed trials (all six periods), the reported mean differences in FIQ scores comparing treatment groups ranged from 31.3 points in favor of AMT + FL (i.e., a 31.3-point decline in FIQ score with this treatment) to 13.6 points in favor of control (i.e., a 13.6-point increase in FIQ score with treatment compared with AMT only). Standard errors for these differences ranged from 1.3 to 21.7, and the size and direction of these mean differences did not correlate directly with the extent of within-patient variation across individuals. Lack of evidence of carryover or time trend supported the assumption of measure independence from time and treatment order [5].

Linear mixed models provide great flexibility for meta-analysis, and we investigated a variety of regression covariates and covariance structures as part of our model building. No added covariates, including practice setting and trial period, contributed significantly to the model with treatment effects (data not shown). So, henceforth, we focus on the treatment effects of AMT and AMT + FL.

Table 1 presents the results of various mixed models using random and fixed effects for the regression intercept and treatment effect as well as unstructured, autoregressive, common, and separate covariance models. We have ordered

Table 1

Results of a representative selection of the assessed mixed models using the treatment parameters and varied assumptions as noted

<table>
<thead>
<tr>
<th>Model</th>
<th>Intercept</th>
<th>Treatment</th>
<th>Patients’ variances</th>
<th>Within-patient variance</th>
<th>Treatment effect</th>
<th>Standard error</th>
<th>Number of parameters</th>
<th>Bayesian information criterion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Random</td>
<td>Fixed</td>
<td>Equal</td>
<td>Single/unstructured</td>
<td>−6.83</td>
<td>2.00</td>
<td>4</td>
<td>1,722</td>
</tr>
<tr>
<td>2</td>
<td>Random</td>
<td>Fixed</td>
<td>Equal</td>
<td>Single/autoregressive</td>
<td>−6.76</td>
<td>2.02</td>
<td>5</td>
<td>1,727</td>
</tr>
<tr>
<td>3</td>
<td>Random</td>
<td>Fixed</td>
<td>Equal</td>
<td>Separate by treatment/ uncorrelated</td>
<td>−6.83</td>
<td>2.00</td>
<td>5</td>
<td>1,727</td>
</tr>
<tr>
<td>4</td>
<td>Random</td>
<td>Random</td>
<td>Equal</td>
<td>Single/uncorrelated</td>
<td>−6.83</td>
<td>2.00</td>
<td>6</td>
<td>1,736</td>
</tr>
<tr>
<td>5</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Equal</td>
<td>Single/uncorrelated</td>
<td>−6.83</td>
<td>2.41</td>
<td>3</td>
<td>1,755</td>
</tr>
<tr>
<td>6</td>
<td>Random</td>
<td>Fixed</td>
<td>Equal</td>
<td>Unstructured</td>
<td>−7.61</td>
<td>1.95</td>
<td>19</td>
<td>1,789</td>
</tr>
<tr>
<td>7</td>
<td>Random</td>
<td>Fixed</td>
<td>Equal</td>
<td>Unstructured</td>
<td></td>
<td></td>
<td>18</td>
<td>1,797</td>
</tr>
<tr>
<td>8</td>
<td>Random</td>
<td>Random</td>
<td>Equal</td>
<td>Unstructured</td>
<td>−7.60</td>
<td>1.95</td>
<td>21</td>
<td>1,800</td>
</tr>
<tr>
<td>9</td>
<td>Random</td>
<td>Fixed</td>
<td>Unequal</td>
<td>Single/uncorrelated</td>
<td>−5.43</td>
<td>1.48</td>
<td>37</td>
<td>1,848</td>
</tr>
<tr>
<td>10</td>
<td>Random</td>
<td>Random</td>
<td>Unequal</td>
<td>Single/uncorrelated</td>
<td>−5.18</td>
<td>1.70</td>
<td>39</td>
<td>1,856</td>
</tr>
<tr>
<td>11</td>
<td>Fixed</td>
<td>Random</td>
<td>Unequal</td>
<td>Single/uncorrelated</td>
<td>−4.03</td>
<td>1.87</td>
<td>37</td>
<td>1,863</td>
</tr>
<tr>
<td>12</td>
<td>Fixed</td>
<td>Fixed</td>
<td>Equal</td>
<td>Unstructured</td>
<td>−4.3</td>
<td>5.1</td>
<td>39</td>
<td>1,882</td>
</tr>
</tbody>
</table>

Models are listed according to the Bayesian information criterion ([22]), with the best model having the lowest numeric value.

These results suggest that:

1. Random intercepts (scores on AMT alone) fit better than fixed intercepts (model 1 vs. model 5, model 6 vs. model 12).
2. Fixed treatment effects fit as well or better than random treatment effects (models 1 vs. 4; models 6 vs. 8, models 9 vs. 10) under a variety of covariance structures.
3. Among the fixed treatment effects models, the common-variance structure (model 1) fits best (model 1 vs. 2, 3, 6, 9, and 12). The addition of an autoregressive term for correlation among time periods (model 2) and an additional variance component for separate variances by treatment group (model 3) do not improve model fit enough to overcome the added parameter. The use of an unstructured covariance matrix (model 6) adds many parameters to the model and has a much worse fit than the simpler covariance structures.
4. The fixed treatment itself is highly significant (model 6 vs. model 7).

Table 2 compares treatment effect estimates from summary and IPD meta-analyses and Bayesian hierarchical models, each with assumptions of fixed and random treatment effects and with common and separate variances for each individual. We used two data sets: balanced trial results from the 34 individuals who completed all six periods (section A) and unbalanced trial results from the 46 individuals who finished at least the first two periods (section B).

Results from the 34 balanced (completed) trials using a common within-patient variance are similar, showing roughly a 7-point advantage (decrease in FIQ score) with treatment, under both fixed- and random-effect assumptions, using summary or individual patient data. Estimates under a separate variance assumption are smaller. Of note, the summary data meta-analysis estimate under the fixed-effect treatment assumption with separate variances is very small and is the only one of all the complex model estimates not to differ significantly from zero. This anomalous result occurs because of a single patient who had an extremely precise estimate of no treatment benefit. The small variance received a large weight in the fixed-effect analysis. Omission of this outlier returned the estimate −6.86, in line with the other models. Very precise treatment effect estimates might result from a trial in which outcomes differ by treatment but where the measurements are highly consistent within each treatment. However, because measurement variation stems from both individual baseline measurements and the treatment effect, measures might likely be more consistent when there is no treatment effect. Using a separate-variance fixed-effect model might then bias the resulting mean estimates toward null. Thus, a major advantage of the random-effect model is its robustness to such outliers.

As expected, the results using the Bayesian hierarchical model with uniform (noninformative) priors for the treatment effect behaved much like the non-Bayesian models. Because the prior treatment mean (8.0) was larger than the treatment mean observed in our study, the posterior means incorporating the informative priors are larger. The posterior standard errors are smaller, reflecting the additional information introduced by the priors.

Assuming a common within-patient variance, results are not sensitive to the choice of variance priors. With unequal between-patient variances, however, results are very sensitive to the prior distributions placed on the within-patient variance (but not the between-patient variances). This is not surprising, because each trial has limited within-patient information (six measures), but when pooled, there are enough trials to provide information on a common within-patient variance and on the between-patient variances. A similar impact is seen using a stronger (more informative) prior. Not surprisingly, removing the “outlier” (i.e., reducing heterogeneity) from the random-effect models yields results closer to those of the fixed-effect models.

Figure 1 illustrates the various prior and posterior distributions (for the parameters in a model with a common within-patient variance, random intercepts, but fixed treatment
Table 2
Mean treatment effect estimates\(^a\) obtained using various analytic models with data from the 34 completed (six periods) trials and from the 46 trials with at least one period completed on each treatment

<table>
<thead>
<tr>
<th>Between-patient variance</th>
<th>Treatment effect prior</th>
<th>Variance priors</th>
<th>Mean difference ± standard error(^b)</th>
<th>Mean difference ± standard error(^b)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Between patients</td>
<td>Within patients</td>
<td>A: 6-Period completers (N = 34)—equal number of periods</td>
</tr>
<tr>
<td>Common</td>
<td>Parallel design(^e)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>AB/BA crossover(^d)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Averaged crossover(^f)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Summary meta-analysis</td>
<td>—</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Individual patient data: likelihood</td>
<td>Noninformative(^i)</td>
<td>Uniform(^h)</td>
<td>Uniform(^h)</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Weak(^g)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Strong(^g)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Informative(^g)</td>
<td>Uniform(^h)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Weak(^f)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Strong(^f)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Individual patient data: Bayesian</td>
<td>Noninformative(^d)</td>
<td>Uniform(^h)</td>
<td>Uniform(^h)</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Weak(^d)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Strong(^d)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td>Informative(^d)</td>
<td>Uniform(^h)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Weak(^d)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
<tr>
<td></td>
<td>Strong(^d)</td>
<td>—</td>
<td>—</td>
<td>—</td>
</tr>
</tbody>
</table>

Abbreviations: FIQ, Fibromyalgia Impact Questionnaire; AMT, amitriptyline; FL, fluoxetine; IG, inverse gamma; SE, standard error.

\(^a\) Mean difference = difference in FIQ scores on the combination (AMT + FL) treatment minus the control (AMT) treatment. A lower FIQ score reflects better condition status, and a negative mean difference reflects greater improvement on the combination treatment.

\(^b\) Confidence intervals = mean estimates ± (SE multiplied by 1.96).

\(^c\) t-test using results of first periods only.

\(^d\) Paired t-test of outcomes on AMT + FL minus AMT from first pair of periods.

\(^e\) Paired t-test of each trial’s averaged outcome on AMT + FL minus averaged outcome on AMT from all completed periods from each trial.

\(^f\) Using posterior distribution of variance parameters as a new prior $t_2^2 \sim \text{IG}(9.99,970.59)$, $t_2^2 \sim \text{IG}(2.742,29.19)$ and $\sigma^2 \sim \text{IG}(83.03,17,003.35)$.

\(^g\) Omitting outlier, estimate is $-6.86 \pm 1.64 (P = 0.0001)$; omitting the outlier from the Bayesian analyses of the balanced trials yields outcomes similar to the equal-variance models.

\(^h\) If assuming unequal variances, test of the heterogeneity of the means supports using a random-effect model ($P = 0.03$).

\(^i\) Not enough data to calculate individual patient variances.

\(^j\) Using trials with 3 or more periods (N = 41): $-3.7 \pm 1.1$ and $-4.6 \pm 1.4$, respectively.
The empirical posterior distributions, for the intercept ($a$) and slope ($\beta$), are well approximated by normal distributions matched with the parameters’ respective posterior means and standard deviations (estimated from the Markov chain Monte Carlo model). These posteriors are more precise than their respective prior distributions, but their rough agreement reflects the information on the means available from the prior published study. The empirical posterior distributions of the between-patient variances ($\tau_a$ and $\tau_b$) and the within-patient variance ($\sigma^2$) are well estimated by IG distributions but with larger scale (corresponding roughly to the degrees of freedom) and shape (a sum of squares about the mean) parameters than the initially chosen prior distributions. These reflect the additional information available from the data, not available from the previous trial.

The right-hand side of Table 2 (section B) presents results using the various models with data from the 46 participants who completed at least two treatment periods (12 noncompleters finished fewer than six periods). The estimates and standard errors using these unbalanced trial designs indicate that the additional information improved precision and also reduced the average treatment effect.

The latter suggests that noncompleters were more likely to have small or no treatment effects and that was the most common reason given for dropping out [5]. The extreme increased precision seen with likelihood analysis of IPD likely reflects the incorrect estimation of precision when no repeated measures are available. If only trials with more than two periods are used, variance estimates are increased (see Table 2 footnotes, n).

In Table 2, we also present the results of analyses that mirror more standard parallel and crossover trial designs. Using only results from the first trial period, analyses mirror a parallel-group design. The first pair of periods and the averaged treatment outcomes from the $N$-of-1 trials were analyzed as a standard and averaged AB/BA crossover trial designs, respectively. Compared with other models, the average treatment effects estimated using the single-period parallel design were somewhat higher ($-8.1 \pm 5.6$) using data from the 34 completed $N$-of-1 trials and somewhat lower ($-5.3 \pm 4.6$) using data from the 46 trials with at least two periods completed. Standard errors were higher for both, and neither one was statistically significant (0.16 and 0.26 using 34 and 46 trials, respectively). The estimates...
using the two-period crossover design were \(-8.0 \pm 3.7\) and 
\(-6.7 \pm 3.3\), using the 34 and 46 trials, respectively, and 
both were statistically significant \((P = 0.04\) and 0.05, re-
spectively). As expected, using the second-period measure-
ments substantially improved the estimates’ precision, and 
as shown in Table 2, adding repeated measures further im-
proved precision.

In the case of the averaged crossover analysis, the esti-
mated effect using data from the 34 balanced trials was 
similar to the results from the other non-Bayesian com-
mon-variance models, although the standard error was 
slightly different. In the averaged crossover analysis, the ef-
ef ect estimate and precision were derived by equally weight-
ing the treatment differences from each trial, ignoring the 
number of periods each had completed. In effect, this 
means that the between-study variance is accounted for 
but not the within-study variances. In contrast, the fixed-
effect common-variance estimate uses the within-study var-
iances but not the between-study variance. In both analyses, 
then, total variance is underestimated, because one of the 
two variance components is ignored.

A question raised then relates to the trade-off between 
studying many patients for a few time periods or few pa-
tients for many time periods. Table 3 shows the relative ef-

ciency achieved by varying the number of pairs of 
treatment periods and the number of patients enrolled for 
three different combinations of within-patient and between-
patient variances in a random-effect model. Table 3 (A) 
uses the estimates from our data (200 and 100, respec-
tively), whereas Table 3 (B and C) show results when the 
within-patient or between-patient variance is reduced. We 

 included some extreme hypothetical designs (e.g., an im-
practical trial with 100 treatment pairs) to illustrate that the 
participant—period trade-off relates not only to analytic 
efficiency but, perhaps, even more to feasibility and practi-
cal considerations, including recruitment and retention. As 
described in our implementation study [5], this ultimately 
shapes the available data. As expected, precision increases 
with longer follow-up (more repeated measures) and with 
greater enrollment. However, the number of patients and 
the number of periods are not of equal importance. For ex-
ample, with our study’s variances (between-patient vari-
ance of 100 and within-patient variance of 200), taking 
10 pairs of measurements on 25 patients (500 measure-
ments) is only about twice as efficient as taking one pair of 
measurements on 46 patients (92 measurements). In this 
case (Table 3 [A]), doubling the number of patients doubles 
efficiency, but doubling the number of pairs increases effi-
ciency by a fraction that decreases as the number of pairs 
increases. After the first several repeats, therefore, addi-
tional measurements add little to precision.

This relative efficiency changes, however, with different 
amounts of between- and within-patient variabilities. Table 
3 (B) demonstrates that additional measurements on indi-

dividual patients remain valuable when the between-patient 
variability is small compared with the within-patient 


variability. Thus, 10 pairs of measurements on 70 patients 
are nearly as valuable as one pair on 60 patients. Con-
versely, when the within-patient variance is small compared 
with the between-patient variance (Table 3 [C]), additional 
measurements on patients add little, because the action is 
occuring between, rather than within, patients.

4. Discussion

In this article, we have explored analytic approaches for 
combining N-of-1 trials’ data to estimate a treatment effect 
for the population. N-of-1 trials not only provide information 
about comparative response but also about the variabil-
y of an individual’s responses to each treatment. A usual 
premise in undertaking N-of-1 trials is that individuals dif-
fer in their responses to treatments and that repeated mea-


surements help to determine the comparative effectiveness 
of the interventions for an individual more accurately. In 
typical population studies, on the other hand, we assume 
that participants are similar and pool their results to esti-


mate an average effect that can be applied to future patients 
in the population. Thus, in combining N-of-1 trials, we 
struggle to balance individuals’ differences with presumed 
similarities in deciding how best to incorporate the 
within-patient information into a representative estimate 
of the treatment effect for the population.

We considered and tested models with varied frame-


works and assumptions. Paralleling the meta-analysis 
of multiperson trials, each of the N-of-1 studies could be 
viewed as a separate trial and their summary data combined 
[13]. Alternatively, the N-of-1 trials might be viewed as 
IPD and their results handled as independent or correlated 
repeated measures [16,17]. Both approaches support incor-
porating prior information through Bayesian hierarchical 
models. Also, lack of any detected trends within patients 
over time supports the N-of-1 trial assumption of indepen-
dent measurements.

We found that under a variety of within-patient covari-


ance structures, the data supported a fixed treatment effect 
rather than a random treatment effect. This implies that the 
effect of added FL is the same for all patients, as is 
generally assumed in a clinical trial. In this respect, the 
meta-analysis of our N-of-1 trials differs from many 
meta-analyses of clinical trials that require random effects 
to model differences in treatments across studies. This is, 
perhaps, not surprising, as the N-of-1 trials we used had 
a common protocol. This assumption may not hold when 
combining N-of-1 trials that have different protocols or if 
the therapies are unrelated (e.g., not a common treatment 
[AMT] with or without an addition [FL]). Patient outcomes 
did differ, however, as evidenced by the need for random 
intercepts, which reflects the different FIQ levels on 
AMT across patients. The lack of any additional explana-
tory power by patient-level covariates to explain the 


between-patient variation reinforces the need for larger
numbers of patients to explore the heterogeneity of patient response. Further study (e.g., simulations) is needed to assess the impact of repeated measures in facilitating discovery of treatment interactions with risk factors.

The best-fitting model used a common within-patient variance with zero correlation between measurements. As with the fixed treatment effect, a common variance is a standard assumption for estimating an overall population effect in many parallel and crossover trial designs. However, because an N-of-1 design with multiple crossovers intends to measure the effect of treatment on an individual, we might anticipate that N-of-1 data would support estimation using individual-specific variances. Our analysis, however, did not find any heteroscedasticity of variances, most likely because six crossovers were too few to estimate these variances well enough. In other words, the data provided insufficient evidence to reject the simpler common-variance assumption. An N-of-1 trial can be designed with more than six crossovers, but in practice, participants may not tolerate additional treatment switches, especially, if there are clinically appreciable differences in the treatments’ effectiveness. As such, for our series and in many others, more complex covariance structures may not be supported. The lack of correlation between period responses in the individuals’ trials also suggests that the time intervals between measurements provided sufficient washout periods to eliminate any correlations.

As expected, the use of prior information modified our knowledge about the value and precision of estimates of model parameters. Because the average treatment effect size in our study was smaller than that in the previous crossover trial (mean difference \(-8.0 [23]\)), the posterior effects were reduced compared with the prior ones but increased compared with those from our present study’s data. These posterior estimates were also much more precise than the estimates based only on our trials’ data. As more N-of-1 trials accrue, this precision only increases. The use of information from other N-of-1 trials and other clinical trials will then strengthen the conclusions drawn about the average treatment effect. Practically, a sequential approach, similar to cumulative meta-analysis [26], which builds these prior distributions from a combination of clinical experience and pilot studies within-patient variation, would seem recommended.

The multilevel structure accounts for both within- and between-patient variations and, thereby, gives more accurate predictions than those by either analyzing each individual trial separately or pooling all the results together with a common slope and intercept [27]. However, it is important to keep in mind that results are sensitive to assumptions about the prior distribution of the variances. Future experience with N-of-1 studies may allow the construction of more informative variance priors. These may help make posteriors more robust to outliers as well. Posterior variances from previous studies may be useful, in this regard. Results with our data-derived priors demonstrated the particular sensitivity of the random-effect models to within-patient variance information. The posteriors from our current study might be used with the next series of N-of-1 trials for fibromyalgia therapies using the FIQ.

In general, additional data increased efficiency. The N-of-1 multi-crossover design increased precision compared with a single-period design or a two-period AB/BA design. Using the averaged crossover analysis, the results reflect the advantage of the repeated measures when the trials have balanced designs. Less clear is the benefit of this averaging approach for combining trials with varied repeated measures. Inclusion of partially completed trials (the 12 additional trials with at least two completed treatment periods) generally increased precision but did not take into account any potential bias introduced by the missing periods and was susceptible to the effect of outliers. Although incorporating partially completed trials is important, analysis must carefully consider potential biases from incomplete data and the increased analytic complexities.

In combining N-of-1 trials, we also need to keep in mind the broader research considerations and our reasons for combining these trials. One can use the combined N-of-1 approach to aggregate data from already completed trials,
each of which presumably was undertaken solely for an individual’s treatment management. This enables the estimation of a population effect and better estimation of individual’s effects by borrowing strength from other patients. In this case, the $N$-of-1 trial design’s increased cost (time and materials) may be justified by the individual benefits that originally supported the use of these trials for personal treatment management (e.g., greater evaluation validity, outcome precision, and patient involvement in self-care). If we are considering designing a prospective study, however, we want to consider what, if any, are the benefits of using a series of $N$-of-1 trials for obtaining an overall population estimate rather than using a more standard trial design.

Although more treatment periods and informative prior distributions increase model estimate precision, they cannot substitute for larger numbers of patients if we seek increased knowledge about between-patient heterogeneity. That said, particularly when dealing with fast-acting treatments with short-lasting effects, allocating resources to more visits per patient rather than recruiting more patients may enable $N$-of-1 designs to be more efficient than single-period designs. As Table 3 indicates, additional periods are most useful when within-patient variances are large compared with between-patient variances.

Nonstatistical considerations may also motivate the use of the combined $N$-of-1 design. These include improving the evaluation of “real world” effectiveness, promoting head-to-head comparative effectiveness studies of already approved or alternative therapies that might not otherwise be studied, increasing trial participant diversity through practice-based research and providing personalized comparative effectiveness assessments to participants. Considerations that extend beyond data analysis may impact cost—benefit evaluations of this research approach. For example, preliminary reports suggest that making individualized outcome data available by using $N$-of-1 trials in population studies may enhance participation in clinical trials [5,12]. Individual and practical implementation benefits coupled with the potential to extend clinical research into routine clinical care might favor this approach over a standard multiperiod crossover design, in which participants’ results are not individually assessed and used.

A combined $N$-of-1 approach to estimate population effects comes with a large caveat regarding generalizability of the resulting population estimate. As with other population research, outcome generalizability depends on how representative trial patients are of the overall population. In our implementation study [5], although we sought broad inclusion, those ultimately enrolled represented only a narrow subset of treatment-naive patients with FMS. As in all clinical trials, patient selection impacts outcome, and this impact may be greater in the combined $N$-of-1 design if resource allocation to repeated measurements limits the range of patients studied.

We conclude by noting that the combined $N$-of-1 trial design permits a sequential approach in estimating the average treatment effectiveness using patients recruited to a study primarily focused on individual effectiveness. The multiple measurements collected on each patient enable the estimation of patient-specific effects and their precision and also enable the researcher to balance individual and population effectiveness objectives. With consideration of the resulting model complexities, the use of these designs may enhance research capacity.

Acknowledgments

We thank Dr. Ingram Olkin for his helpful discussions. We also thank the journal reviewers for their very helpful suggestions for improving this manuscript and, in particular, for suggesting the addition of the averaged crossover analysis to our comparison.

Funding and support: This research was funded in part by a grant # RO1 AR45416 (PI: Dr. Zucker) from the National Institute of Arthritis and Musculoskeletal and Skin Diseases.

Role of the sponsors: The National Institute of Arthritis and Musculoskeletal and Skin Diseases did not participate in the design and conduct of this study; in the collection, analysis, or interpretation of the data; or in the preparation, review, or approval of this article.

References