ABSTRACT. Objective. Applying population research to individual treatment requires understanding the connections between patient-specific characteristics, population-based studies, and treatment responses. Conducting practice-based research using individual-focused (N-of-1) trials may aid this process. We combined N-of-1 trials to compare fibromyalgia therapies and to assess the feasibility and outcomes of this approach for practice-based effectiveness research.

Methods. Community- and center-based rheumatologists enrolled patients with fibromyalgia syndrome in randomized, double-blind, multi-crossover, N-of-1 trials comparing amitriptyline and the combination amitriptyline and fluoxetine. Fibromyalgia Impact Questionnaire outcomes were used for the individuals’ treatment and combined across patients for sample-based analyses. Outcomes were compared with results from more standard trial designs.

Results. Eight rheumatologists enrolled 58 patients in N-of-1 trials. Most physicians and patients had not previously participated in clinical trials. Using several analytic methods, the pooled results showed a better outcome score (mean difference: −6.1 ± 2.0 to −8.0 ± 3.7 points) in patients taking combination therapy. These population results are similar to published outcomes from a more traditional crossover trial. Neither practice type nor patient characteristics were significantly associated with the observed treatment-effect variation. Most participants, irrespective of selected treatment, felt their individual N-of-1 trials were helpful.

Conclusion. Implementation of the combined N-of-1 methodology is feasible in rheumatology practice and results confirm greater fibromyalgia improvement with combination therapy. This research approach broadens participation, although our trials’ specifics likely influenced enrollment eligibility. In addition to individual benefits, combining N-of-1 trial data provides population research benefits. This patient-focused approach should be further explored to bridge research and practice. (J Rheumatol 2006;33:2069–77)

Key Indexing Terms:
N-of-1 EFFECTIVENESS
FIBROMYALGIA AMITRIPTYLINE PRACTICE-BASED FLUOXETINE

An N-of-1 trial is a multi-crossover study in which an individual compares her/his responses to, generally 2, treatments and assesses the variability in these responses. The N-of-1 crossover design is particularly applicable to the study of symptomatic treatments for chronic conditions such as fibromyalgia syndrome (FM). FM is a common cause of widespread, chronic pain and, despite increased research, effective therapies remain limited. Assessing FM therapies is particularly challenging since patients have multiple symptoms that can vary greatly. As such, a crossover trial design in which patients serve as their own comparison is particularly useful. In 1996, the combination of amitriptyline and fluoxetine (AMT+FL) was shown to provide a benefit to patients with FM equal to the sum of the treatment effects of each drug alone. However, only 63% of that study’s patients showed clinically significant improvements with this treatment. Treatment response variation seen in this and other FM trials may relate to patient subtypes.

This individual subject variability poses a challenge for research and for bridging population research findings and individual patient treatments. Partnering academic researchers and community-based practitioners for practice-based research could help to diversify research participants and

Zucker, et al: Combining n-of-1 trials of FM therapies

Personal non-commercial use only. The Journal of Rheumatology Copyright © 2006. All rights reserved.
improve our understanding of the connections between individual patient characteristics and population-based treatment responses. Using study designs that can mesh individual treatment management and research may be helpful.

Research has supported the usefulness of N-of-1 trials for individual treatment decision-making and has recognized its parallels to clinical practice. More recently, with growing interest in evidence-based, individualized medicine, studies have explored the cost-effectiveness of using more rigorous N-of-1 trials for effective clinical decision-making. Using N-of-1 trials may lead to outcome improvements and reduce unnecessary medication costs. Less studied, however, is the potential for combining the individually beneficial N-of-1 trials’ results for population-based research. This can integrate research and practice and, if carried out in community practices, might broaden applicability.

We chose therefore to implement the combined N-of-1 practice-based research approach to compare responses to AMT and the combination therapy (AMT+FL) among patients with FM at community rheumatology practices and at a FM referral center. We used individuals’ results to guide patient-physician treatment decisions and to obtain comparative effectiveness estimates for the population. These estimates were then compared with published results from a standard trial design. In addition, we explored similarities and differences in the combined trials’ outcomes from different practice settings and among patients with specific characteristics that might influence treatment responses. This study allows us to draw preliminary conclusions about using the combined N-of-1 design to link research and clinical practice in individuals and in patient populations.

MATERIALS AND METHODS

Physician recruitment. Rheumatologists in community practices were sought (enrollment target 6–10 physicians) to conduct N-of-1 trials in their practices. Lists of community-based male and female rheumatologists in the Boston area were randomly ordered. Invitations, in sequential batches of up to 12 invitations/batch, were mailed and followed up with a telephone call. We allowed 2–3 weeks for responses before mailing the next batch of invitations. Remuneration to participating physicians was $150.00 for each patient visit.

Patient recruitment. Persons between 18 and 60 years of age, meeting American College of Rheumatology criteria for FM, without comorbid conditions that might confuse FM diagnosis or assessments, with no contraindications to study medications, and with a score not greater than 18 on the Beck Depression Inventory (BDI) scale were eligible for enrollment. Women of childbearing potential were required to have no plan to conceive and to agree to use appropriate contraception. Participants agreed to stop their current FM medications for at least one week prior to starting the trial and were not allowed to use analgesics other than acetaminophen during the study. These criteria were intentionally similar to those of the earlier study to allow direct comparisons of the studies’ results for the 2 therapies (AMT and AMT+FL).

Modes of patient recruitment varied among the physicians’ practices and included electronic systems to identify and contact potentially eligible patients, recruitment as patients with FM were encountered in practices, and mailings to patients with FM. Due to unexpected low patient eligibility and accrual within the practices, the rheumatologists subsequently sent letters to their referring primary-care physicians, informing them of the study for possible referrals of eligible patients. Additionally, advertisements were placed in local newspapers. Respondents were screened and if appropriate, referred to the geographically most convenient participating community-based rheumatologist. She confirmed study eligibility and obtained informed consent for participation. No financial incentive was offered to patients for participating in an N-of-1 trial.

Study design. Each N-of-1 trial (Figure 1) had six 6-week intervention periods (3 paired sets). In each set, patients received treatment with AMT (placebo in the morning and 25 mg AMT at night) and with AMT+FL (20 mg FL in the morning and 25 mg AMT at night). A central pharmacy prepared the blinded, random-ordered treatment kits and block-randomized the first treatment sets, so for every 20 kits prepared, 10 started treatment with AMT and 10 started with AMT+FL. This trial design used the same medication dosages, treatment period length, and outcome measures as the published study to enable direct comparisons.

Baseline evaluations included: depression screening, laboratory tests (liver functions, complete blood cell assessment, electrocardiogram, a pregnancy test where applicable), and baseline outcome assessments. At the end of each 6-week treatment period, patients were evaluated and received their next packet of medications to start. The Goldenberg study had demonstrated that Fibromyalgia Impact Questionnaire (FIQ) scores returned to baseline 2 weeks after stopping these medications. At each visit, participants were asked about any other medications taken and any adverse symptoms experienced.

The main outcome measure was the score on the FIQ, a validated, multi-part quality of life assessment tool scored from 0 (best) to 100 (worst score). Additional outcomes measured included: patient-assessed visual analog scales (VAS) specifically gauging global well-being, pain, sleep, fatigue, and feeling refreshed upon awakening, from 0 mm (best) to 100 mm (worst), as well as 2 physician-assessed VAS of patient’s global well-being and patient’s stress. Rheumatologists also performed manual examination of patients’ 18 FM tender points and the quantified responses were summed together for an overall tender point examination (TPE) score (range 0–36). Study outcomes were measured at baseline, at the end of each 6-week treatment period, and in the 3-month follow-up visit.

Individual trials’ outcomes. At the end of a completed N-of-1 trial (6 periods) individuals’ FIQ scores on the 2 therapies were compared. Only the FIQ results, analyzed and displayed both graphically and quantitatively, were provided to the physician and patient for use in choosing a therapy that the physician then prescribed. At the end of our overall study, however, we evaluated the correlations between other outcome measures and individuals’ FIQ scores in response to the therapies. Participants completed questionnaires about their N-of-1 trial perceptions and experiences at baseline, upon trial completion, and 3 months later.

Safety. Participants encountering any difficulties or wanting to withdraw from the study could contact their physicians at any time and be withdrawn. The AMT and FL interaction was assessed by measuring serum AMT and nortriptyline (AMT/NOR) levels at the end of the first and second periods (once while taking each treatment). Any combined level exceeding an average antidepressant therapeutic level (200 µg/l) was considered significant. Adverse events were reviewed by the principal investigator and reported to the Institutional Review Board (IRB) and to the study’s safety officer. Our study was approved by the IRB at New England Medical Center, Newton-Wellesley Hospital, and Sturdy Memorial Hospital, Attleboro, MA, USA, and there was an independent external safety monitor.

Combined N-of-1 population outcomes. Results from the N-of-1 trials were combined to obtain population estimates of treatment effectiveness. Estimates were calculated using data from all participants who finished at least 2 treatment periods (one while taking each therapy) and, separately, using data from participants completing all 6 of their N-of-1 trials’ periods. Analytic models
with and without incorporating prior treatment-effect information from the Goldenberg study were used (see Bayesian analyses below).

Enrollment site (center or community) and selected patient characteristics (age, sex, body mass index, height, weight, weight change, symptom duration, employment status, marital status, weekly exercise, drug use, presence of irritable bowel, facial pain and/or migraine, baseline outcome scores, composite baseline severity, belief of treatment efficacy, patient anxiety level, prior FM therapy) were separately tested for significant association with treatment effectiveness. Carryover effect and time trend, both concerns when using crossover and repeated measures designs, were assessed across individuals.

**Combined N-of-1 outcome comparisons.** The combined N-of-1 population estimates were compared with outcomes from more typical crossover trial designs in which participants try each treatment only once. The first measurements on each therapy from each N-of-1 trial were compared as in a single (AB/BA) crossover trial. These outcomes (using FIQ scores and the other measures) were then directly compared with the analogous (AMT and AMT+FL) outcomes from the published crossover trial.

**Statistical methods.** Classic individual N-of-1 trial analyses comparing FIQ scores on the 2 therapies used a 2-sided t-test assuming equal variances for the 2 therapies.

Intercorrelations of FIQ scores and the other outcome measures were determined using Spearman (rank) correlation coefficients. Median correlations across the individuals completing at least one period taking each therapy were calculated.

Carryover and time trend significance were tested using random-effects regression models that included treatment pattern variables and time together with a time-by-treatment interaction term, respectively.

Bayesian analyses combining N-of-1 trials employed a 2-level random-effects model to describe the posterior distributions of treatment effectiveness. For these analyses we assumed a common within-patient variance. More complex variance structures did not improve model performance. Analyses used both noninformative and, separately, informative priors derived from published trial results (see details in Appendix).

**RESULTS**

Recruitment and enrollment. Seven community-based rheumatologists agreed to participate in response to 69 invitations sent (10% response rate). One of these rheumatologists was actively involved with practice-based research at the time of recruitment. The others reported some research experience during their subspecialty training. No participating physician had prior N-of-1 trial experience.

Fifty-eight patients enrolled in N-of-1 trials from November 2000 through February 2003. Fifty-two were enrolled from the physicians’ patient panels that together with and without incorporating prior treatment-effect information from the Goldenberg study were used (see Bayesian analyses below).

Enrollment site (center or community) and selected patient characteristics (age, sex, body mass index, height, weight, weight change, symptom duration, employment status, marital status, weekly exercise, drug use, presence of irritable bowel, facial pain and/or migraine, baseline outcome scores, composite baseline severity, belief of treatment efficacy, patient anxiety level, prior FM therapy) were separately tested for significant association with treatment effectiveness. Carryover effect and time trend, both concerns when using crossover and repeated measures designs, were assessed across individuals.

Combined N-of-1 results were compared with a more standard crossover design. The first pairs of measures (one on each treatment) across the N-of-1 trials were compared using a paired t-test. These outcomes were compared between groups (center and community, N-of-1, and published results) using an analysis of variance (ANOVA).

Estimated benefits of N-of-1 trials for individuals and for research: Among trial completers who selected either AMT or AMT+FL, we compared the proportion selecting AMT with random AMT selection (probability = 0.5) using a one-sample t-test. Additionally, outcomes of individuals’ N-of-1 based therapy choices (i.e., the mean FIQ score from the 3 periods taking their selected treatment) were compared with their similarly determined FIQ scores for empiric use of a single agent (AMT) and for empiric combination therapy. For these 3 choice strategies across patients, the overall mean FIQ scores were compared using an ANOVA and subsequent pairwise comparisons. For the population estimates, we determined the relative numbers of patients needed to enroll to attain similar precision using a combined N-of-1 approach as compared with a single crossover (AB/BA) design.

Statistical analyses were carried out using SAS System for Windows 8.2 and S-plus V.6.2. For Bayesian hierarchical model calculations using Markov chain Monte Carlo techniques, we used WinBUGS Version 1.4.

**RESULTS**

Recruitment and enrollment. Seven community-based rheumatologists agreed to participate in response to 69 invitations sent (10% response rate). One of these rheumatologists was actively involved with practice-based research at the time of recruitment. The others reported some research experience during their subspecialty training. No participating physician had prior N-of-1 trial experience.

Fifty-eight patients enrolled in N-of-1 trials from November 2000 through February 2003. Fifty-two were enrolled from the physicians’ patient panels that together with and without incorporating prior treatment-effect information from the Goldenberg study were used (see Bayesian analyses below).

Enrollment site (center or community) and selected patient characteristics (age, sex, body mass index, height, weight, weight change, symptom duration, employment status, marital status, weekly exercise, drug use, presence of irritable bowel, facial pain and/or migraine, baseline outcome scores, composite baseline severity, belief of treatment efficacy, patient anxiety level, prior FM therapy) were separately tested for significant association with treatment effectiveness. Carryover effect and time trend, both concerns when using crossover and repeated measures designs, were assessed across individuals.

**Combined N-of-1 outcome comparisons.** The combined N-of-1 population estimates were compared with outcomes from more typical crossover trial designs in which participants try each treatment only once. The first measurements on each therapy from each N-of-1 trial were compared as in a single (AB/BA) crossover trial. These outcomes (using FIQ scores and the other measures) were then directly compared with the analogous (AMT and AMT+FL) outcomes from the published crossover trial.

**Statistical methods.** Classic individual N-of-1 trial analyses comparing FIQ scores on the 2 therapies used a 2-sided t-test assuming equal variances for the 2 therapies.

Intercorrelations of FIQ scores and the other outcome measures were determined using Spearman (rank) correlation coefficients. Median correlations across the individuals completing at least one period taking each therapy were calculated.

Carryover and time trend significance were tested using random-effects regression models that included treatment pattern variables and time together with a time-by-treatment interaction term, respectively.

Bayesian analyses combining N-of-1 trials employed a 2-level random-effects model to describe the posterior distributions of treatment effectiveness. For these analyses we assumed a common within-patient variance. More complex variance structures did not improve model performance. Analyses used both noninformative and, separately, informative priors derived from published trial results (see details in Appendix).

**RESULTS**

Recruitment and enrollment. Seven community-based rheumatologists agreed to participate in response to 69 invitations sent (10% response rate). One of these rheumatologists was actively involved with practice-based research at the time of recruitment. The others reported some research experience during their subspecialty training. No participating physician had prior N-of-1 trial experience.

Fifty-eight patients enrolled in N-of-1 trials from November 2000 through February 2003. Fifty-two were enrolled from the physicians’ patient panels that together with and without incorporating prior treatment-effect information from the Goldenberg study were used (see Bayesian analyses below).

Enrollment site (center or community) and selected patient characteristics (age, sex, body mass index, height, weight, weight change, symptom duration, employment status, marital status, weekly exercise, drug use, presence of irritable bowel, facial pain and/or migraine, baseline outcome scores, composite baseline severity, belief of treatment efficacy, patient anxiety level, prior FM therapy) were separately tested for significant association with treatment effectiveness. Carryover effect and time trend, both concerns when using crossover and repeated measures designs, were assessed across individuals.

**Combined N-of-1 outcome comparisons.** The combined N-of-1 population estimates were compared with outcomes from more typical crossover trial designs in which participants try each treatment only once. The first measurements on each therapy from each N-of-1 trial were compared as in a single (AB/BA) crossover trial. These outcomes (using FIQ scores and the other measures) were then directly compared with the analogous (AMT and AMT+FL) outcomes from the published crossover trial.

**Statistical methods.** Classic individual N-of-1 trial analyses comparing FIQ scores on the 2 therapies used a 2-sided t-test assuming equal variances for the 2 therapies.

Intercorrelations of FIQ scores and the other outcome measures were determined using Spearman (rank) correlation coefficients. Median correlations across the individuals completing at least one period taking each therapy were calculated.

Carryover and time trend significance were tested using random-effects regression models that included treatment pattern variables and time together with a time-by-treatment interaction term, respectively.

Bayesian analyses combining N-of-1 trials employed a 2-level random-effects model to describe the posterior distributions of treatment effectiveness. For these analyses we assumed a common within-patient variance. More complex variance structures did not improve model performance. Analyses used both noninformative and, separately, informative priors derived from published trial results (see details in Appendix).
included over 2000 patients with active FM. Rheumatologists reported that comorbid depression, prior use of the trial medications, comorbidities confusing FM evaluation (in particular, rheumatoid and osteoarthritis), and lack of interest in participating in a study were the main barriers in recruiting their patients. Of 90 newspaper advertisement responders, 9 (10%) completed the evaluation processes and were found to be eligible for study and 6 enrolled.

Baseline characteristics were similar (Table 1) between patients from the referral center (n = 34) and the community practices (n = 24). These participants were also similar to the published study’s center-based FM cohort (data not shown). Of the N-of-1 trial enrollees, 88% reported no previous research participation. Forty-six participants (79%) completed at least 2 treatment periods (one taking each therapy) and 34 (59%) completed their N-of-1 trials (6 periods). Retention in the trials and reasons for withdrawal are summarized in Table 2. Overall dropout rates were similar between the practice types (p = 0.97).

Safety. AMT and/or NOR were detectable (≥ 25 µg/l) in only 20% of the tested samples. All combined AMT/NOR levels were ≤ 30 µg/l taking AMT and ≤ 70 µg/l taking AMT+FL. Side effects were reported by 38 participants during their trials and led to 8 patient withdrawals (Table 2). The most frequently reported side effects were sedation, headache, dryness, and gastrointestinal-related symptoms/change in bowel habits.

Individual outcomes and feedback. Individuals’ differences in mean FIQ scores while taking AMT+FL compared with AMT are shown in Figure 2 together with the 95% confidence interval (CI). No correlation is noted between the size or direction of the mean differences and the participants’ response variations (shown as the CI). FIQ score correlations with the other outcomes varied across individuals and the median correlations with the other quality of life measures were high: VAS-GLOBAL (0.77), VAS-PAIN (0.77), VAS-SLEEP (0.77), VAS-FATIGUE (0.71), VAS-REFRESHED (0.71), and Global MD-VAS (0.77). The FIQ score was less well-correlated overall with the TPE score (0.46).

Reports of individuals’ FIQ scores and analyses were provided to the 34 trial completers. Among these participants, 22 chose combination therapy (2 modified the dose of AMT due to sleep/wakefulness issues), 8 chose AMT alone (3 in modified dosages), and the remaining 4 patients chose other therapies. Most (85%) selected the therapy with the lower mean FIQ score. In our study, the large majority (over 90%) of trial completers reported their N-of-1 trial was useful and would recommend it to other patients. At the end of their trials, 94% reported they would consider participating in another N-of-1 trial, and 87% held that view after 3 months.

Population outcomes and comparisons with more standard trial designs. Combined N-of-1 trials’ estimates of treatment effects differed somewhat depending on the analytic method used (e.g., all data vs first pair only) and the populations included (2-period vs trial completers), but all supported a significantly greater FIQ score improvement taking combination therapy (Table 3). Compared with a single-crossover design, precision is increased using the Bayesian models.

Table 1. Participants’ characteristics.

<table>
<thead>
<tr>
<th></th>
<th>Center, n = 34</th>
<th>Community, n = 24</th>
<th>p</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sex-female (%)</td>
<td>33 (97)</td>
<td>23 (96)</td>
<td>0.80</td>
</tr>
<tr>
<td>Age, yrs (mean ± SD)</td>
<td>44 ± 10</td>
<td>42 ± 9</td>
<td>0.46</td>
</tr>
<tr>
<td>Ethnic Caucasian (%)</td>
<td>12 (86)*</td>
<td>22 (92)</td>
<td>0.56</td>
</tr>
<tr>
<td>Marital, married (%)</td>
<td>25 (74)</td>
<td>17 (71)</td>
<td>0.82</td>
</tr>
<tr>
<td>Yrs of symptoms, median (q1–q3)</td>
<td>4.5 (3–9)</td>
<td>5 (3–8)*</td>
<td>0.83</td>
</tr>
<tr>
<td>History of facial pain (%)</td>
<td>9 (27)</td>
<td>3 (13)</td>
<td>0.20</td>
</tr>
<tr>
<td>History of migraine (%)</td>
<td>14 (41)</td>
<td>10 (42)</td>
<td>0.97</td>
</tr>
<tr>
<td>History of irritable bowel syndrome (%)</td>
<td>21 (62)</td>
<td>12 (50)</td>
<td>0.37</td>
</tr>
<tr>
<td>Baseline outcome scores</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIQ</td>
<td>58 ± 15</td>
<td>58 ± 15</td>
<td>0.93</td>
</tr>
<tr>
<td>Visual assessment scales</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Patient VAS–global</td>
<td>68 ± 22</td>
<td>68 ± 19</td>
<td>0.99</td>
</tr>
<tr>
<td>Patient VAS–pain</td>
<td>70 ± 19</td>
<td>69 ± 18</td>
<td>0.80</td>
</tr>
<tr>
<td>Patient VAS–sleep</td>
<td>72 ± 24</td>
<td>73 ± 24</td>
<td>0.90</td>
</tr>
<tr>
<td>Patient VAS–fatigue</td>
<td>72 ± 17</td>
<td>79 ± 19</td>
<td>0.15</td>
</tr>
<tr>
<td>Patient VAS–refreshed</td>
<td>74 ± 19</td>
<td>84 ± 17</td>
<td>0.03</td>
</tr>
<tr>
<td>Physician VAS–global</td>
<td>73 ± 20</td>
<td>64 ± 23</td>
<td>0.14</td>
</tr>
<tr>
<td>Tender point examination score</td>
<td>21 ± 5</td>
<td>22 ± 7</td>
<td>0.67</td>
</tr>
<tr>
<td>History of prior treatment</td>
<td>23 (70%)**</td>
<td>17 (71%)</td>
<td>0.93</td>
</tr>
<tr>
<td>Beck Depression Inventory score</td>
<td>12 ± 4</td>
<td>10 ± 5</td>
<td>0.06</td>
</tr>
</tbody>
</table>

* Based on N = 14; † based on N = 21; ** based on N = 33. FIQ: Fibromyalgia Impact Questionnaire, VAS: visual analog scale.
Table 2. Retention in N-of-1 trials, reasons for withdrawal, and reports of side effects.

<table>
<thead>
<tr>
<th>Period</th>
<th>No. Enrolled (%) retained</th>
<th>Number: Reason for Dropout During This Period</th>
<th>Side Effect Reports (No. of Patients reporting Side Effects)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Center</td>
<td>Community</td>
<td>No. Withdrawn</td>
</tr>
<tr>
<td>Baseline</td>
<td>34</td>
<td>24</td>
<td>0</td>
</tr>
<tr>
<td>1</td>
<td>31 (91)</td>
<td>22 (92)</td>
<td>5</td>
</tr>
<tr>
<td>2</td>
<td>28 (82)</td>
<td>18 (75)</td>
<td>7</td>
</tr>
<tr>
<td>3</td>
<td>24 (71)</td>
<td>17 (71)</td>
<td>5</td>
</tr>
<tr>
<td>4</td>
<td>23 (68)</td>
<td>15 (63)</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>21 (62)</td>
<td>14 (58)</td>
<td>3</td>
</tr>
<tr>
<td>6</td>
<td>20 (59)</td>
<td>14 (58)</td>
<td>1</td>
</tr>
<tr>
<td>3 month followup</td>
<td>20 (59)</td>
<td>11 (46)</td>
<td>3</td>
</tr>
</tbody>
</table>

* Thirteen patients reported the same side effect(s) during both AMT and AMT + FL treatments and 6 participants reported the same side effect(s) during more than one period taking the same medication. Thirty-two percent of patients reported more than one side effect taking AMT and 39% reported more than one side effect taking AMT + FL. † All were taking AMT + FL. ** 3 out of 4 were taking AMT + FL. AMT: amitriptyline. FL: fluoxetine.

Figure 2. Individuals’ treatment response variations shown as differences in mean FIQ scores taking AMT+FL and AMT with 95% confidence intervals. Center: center-based practice; comm: community-based practice.
Incorporating the published prior probability of an 8-point FIQ score improvement on AMT+FL into our Bayesian analyses slightly increases the treatment-effect estimates.

We tested for interactions of treatment effect and enrollment site (center vs community) or any of the selected patient characteristics. No statistically significant associations were found in our study population (data not shown). Additionally, no significant interactions between time and treatment effect or between treatment order and treatment effect were identified using random-effects regression models (data not shown).

To directly compare outcomes from N-of-1 trial completers with the 2-period completers, and then with the published crossover trial\(^2\), N-of-1 data were reanalyzed as a standard single-crossover trial (Table 4). Only one measure from each therapy (in other words, data only from the first pair of periods) from the N-of-1 trials was used. In the N-of-1 trials, statistically significant improvements taking AMT+FL compared with AMT were found in FIQ scores (–6.7, \(p = 0.05\)) as well as the VAS-pain scale (–11.1, \(p = 0.02\)) and TPE scores (–3.4, \(p = 0.001\)). Compared with the Goldenberg crossover trial outcomes\(^2\), no statistically significant differences were found in FIQ score differences either using results from all N-of-1 trials (\(p = 0.42\)) or using the subset of N-of-1 trials (\(p = 0.66\)) carried out at the same referral center. Similarly, comparing the other measure outcomes, no significant differences were found (data not shown).

**Table 3. Population treatment effect estimates from combined N-of-1 trials’ results.**

<table>
<thead>
<tr>
<th>Analytic Method</th>
<th>Trial Completers (n = 34)</th>
<th>Probability &gt; 0</th>
<th>2-period Completers (n = 46)</th>
<th>Probability &gt; 0</th>
</tr>
</thead>
<tbody>
<tr>
<td>A. Combining all available treatment period data</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIQ Bayesian methods using only data from this study</td>
<td>–6.8 ± 2.1</td>
<td>99.9%</td>
<td>–6.1 ± 2.0</td>
<td>99.9%</td>
</tr>
<tr>
<td>(noninformative priors)*</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIQ Bayesian methods using information from prior study</td>
<td>–7.1 ± 1.8</td>
<td>100%</td>
<td>–6.6 ± 1.7</td>
<td>100%</td>
</tr>
<tr>
<td>B. Combining only first 2 periods’ data (as in AB/BA trial design)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>FIQ 2 period paired t-test</td>
<td>–8.0 ± 3.7</td>
<td>(p = 0.04)</td>
<td>–6.7 ± 3.3</td>
<td>(p = 0.05)</td>
</tr>
</tbody>
</table>

* See details in Appendix. FIQ: Fibromyalgia Impact Questionnaire. AMT: amitriptyline. FL: fluoxetine.

**Table 4. Population treatment effect comparisons combining only the N-of-1 trials’ first 2 periods’ data as in a standard AB/BA trial design.**

<table>
<thead>
<tr>
<th>Measure</th>
<th>Analytic Method</th>
<th>Trial Completers (n = 34)</th>
<th>(p)</th>
<th>2-period Completers (n = 46)</th>
<th>(p)</th>
</tr>
</thead>
<tbody>
<tr>
<td>FIQ</td>
<td>2-period paired t-test</td>
<td>–8.0 ± 3.7</td>
<td>0.04</td>
<td>–6.7 ± 3.3</td>
<td>0.05</td>
</tr>
<tr>
<td>VAS-global</td>
<td>2-period paired t-test</td>
<td>–8.9 ± 5.0</td>
<td>0.08</td>
<td>–7.0 ± 4.8</td>
<td>0.16</td>
</tr>
<tr>
<td>VAS-sleep</td>
<td>2-period paired t-test</td>
<td>–6.1 ± 6.9</td>
<td>0.97</td>
<td>–2.0 ± 5.5</td>
<td>0.71</td>
</tr>
<tr>
<td>VAS-pain</td>
<td>2-period paired t-test</td>
<td>–12.4 ± 5.0</td>
<td>0.02</td>
<td>–11.1 ± 4.8</td>
<td>0.02</td>
</tr>
<tr>
<td>VAS-awaken</td>
<td>2-period paired t-test</td>
<td>–5.8 ± 5.4</td>
<td>0.28</td>
<td>–3.8 ± 4.6</td>
<td>0.42</td>
</tr>
<tr>
<td>VAS-fatigue</td>
<td>2-period paired t-test</td>
<td>–6.5 ± 5.9</td>
<td>0.28</td>
<td>–4.1 ± 5.5</td>
<td>0.46</td>
</tr>
<tr>
<td>TPE</td>
<td>2-period paired t-test</td>
<td>–3.6 ± 1.2</td>
<td>0.005</td>
<td>–3.4 ± 1.0</td>
<td>0.001</td>
</tr>
<tr>
<td>VAS-global MD</td>
<td>2-period paired t-test</td>
<td>–7.2 ± 6.5</td>
<td>0.27</td>
<td>–8.2 ± 5.6</td>
<td>0.15</td>
</tr>
</tbody>
</table>

FIQ: Fibromyalgia Impact Questionnaire; VAS: visual analog scales; TPE: tender point examination.

**Potential benefits of combining N-of-1 trials.** Using N-of-1 trials, significantly fewer participants than expected by random selection (probability = 0.5) chose treatment with AMT alone (27%; \(p\) for one-sample t-test = 0.007). The mean FIQ outcome for the N-of-1 trial-based treatment choices (mean FIQ = 40.3) was low (improved) compared with the mean FIQ score while taking either single (AMT) medication (44.7; \(p < 0.001\)) or uniform AMT+FL therapy (41.4; \(p = 0.27\)). While the latter difference was not statistically significantly lower, the N-of-1 trials-based outcome was attained with reduced medication usage (27% used AMT only).

The ability to use the data from individually beneficial trials for population research itself provides a potential added benefit of these trials. Additionally, combining N-of-1 trials (with repeated measures) for a population estimate may reduce the number of enrollees required for equivalent estimate precision compared with a standard crossover design. In our analyses (Table 3), the standard error is reduced when all measures are analyzed compared with 2-period analyses (–6.6 ± 1.7 vs –6.7 ± 3.3). Using our resulting within- and between-patient variance estimates (200 and 100, respectively), enrollment of 25 patients with 3 pairs of measures was calculated to give about the same precision as 46 subjects with one pair of measures.

**DISCUSSION**

In addition to comparing FM therapies, our study...
methodologic and feasibility insights about using the combined N-of-1 approach for practice-based effectiveness research. We found that this research design can be used in center- and community-based rheumatology practices. It appears capable of broadening research involvement, with 75% of our physician investigators previously not involved in practice-based research and 88% of patients never having been research participants before this trial.

Our population results showed a significantly better FIQ score outcome on AMT+FL compared with AMT alone. These results concur with the population findings of the earlier standard crossover trial outcomes. In combining N-of-1 trials, the effect size estimates differ somewhat depending on the analytic methods and patient samples used (e.g., trial completers vs all 2-period completers). The hierarchical Bayesian approach allows estimation of both population and individual treatment efficacy and can incorporate prior information into both. Further work is under way to compare the various analytic models and their underlying assumptions and limitations in order to identify the most appropriate approach(es) for combining N-of-1 trials’ results.

The combined N-of-1 population design is an extension of both the standard crossover trial and the repeated measures trial and as such provides information not available in more typical designs. The repeated measures on each treatment within and across individuals allowed us to evaluate how well the FIQ scores and other outcomes correlate in measuring within and across individuals allowed us to evaluate how well the FIQ scores and other outcomes correlate in measuring individual patients further dispel the utility of the more “objective” diagnostic TPE as a valid measure for FM treatment response.

Repeated measures on individuals also add to statistical power, as evidenced in the precision of our population estimates, and can reduce enrollment requirements. The appropriateness of this increased exposure needs to be taken into account in designing studies, while also recognizing other potential limitations that enrolling fewer patients can have in addressing research questions. For example, the ability to identify specific patient characteristics associated with response variation depends on overall enrollment and heterogeneity. Among our participants we were not able to identify significant variables (practice site or patient characteristics) associated with the observed response variation.

Although we sought broad enrollment of patients with FM, we found patient recruitment more daunting than anticipated. We believe the difficulty likely reflects our choice of study medications and the inclusion criteria. Many patients with FM in rheumatologists’ practices had already tried the therapies prior to the study. Trial eligibility was further limited by the exclusion of the substantial numbers of patients with FM who also had depression or comorbid rheumatologic conditions.

Confusing FM evaluation. As a result, our participants were largely drawn from the similarly small numbers of relatively “treatment-naïve” patients with FM in rheumatology practices and were alike. Better understanding of the bases for enrollment limitations is important in order to delineate feasibility effects. Interestingly, once enrolled, a high percentage of the participants (79%) completed at least 2 treatment periods (taking each treatment at least once). Excluding discontinuations due to significant side effects (14%; similar to reported trial discontinuation using these medications) this represents 92% of the remaining participants. We need to further ascertain if the individual and management focus of the combined N-of-1 approach affects participant enrollment and retention and, as a result, outcome generalizability.

Successful clinical trials often depend on participants’ ability to adopt the study processes and interventions as well as to perceive a benefit to their participation. Our participating physicians found the N-of-1 trials feasible within their practices. Most patient-participants felt that their N-of-1 trial was helpful to them, would recommend it to others, and would participate again to compare other potential therapies. Although this information was collected only from those who completed the N-of-1 studies (58% of enrollees), it corroborates other reports of individual satisfaction with N-of-1 trials.

Quantifying the benefits of N-of-1 trials poses a greater challenge and may differ for individuals and for a population. Earlier studies compared outcomes of patients randomized to undertake N-of-1 trials for treatment management with groups that did not. We compared outcomes within the same individuals. The outcomes from their N-of-1-based treatment choices showed an improved (lower) FIQ score compared with their outcomes taking empiric AMT or empiric AMT+FL treatment. Compared with AMT+FL, the N-of-1 reduction in FIQ was not statistically significantly lower (p = 0.27), but was achieved with less drug use (8 of 30 patients used only single-drug treatment). As with the prior group-to-group comparisons, finding only limited outcome benefits given the higher costs of N-of-1 trials may deter these trials’ more widespread use. However, to date, the cost-benefit analyses have not taken the population research potential of combining N-of-1 trials into account. While limited to the study of therapies appropriate for more widespread testing (such as Phase IV studies), integrating population estimate determinations with individual, repeated measures treatment management trials might prove cost-effective compared with the combined costs of standard care and research. Further work is needed to evaluate this.

We recognize several limitations of our study and acknowledge the need to extend this work in light of the lessons learned. For example, to combat the recruitment shortcomings, enrolling patients from primary care practices might be considered in planning future studies. The establishment of practice-research networks could help physician recruit-
ment. More rapid design and funding of practice-based research is needed to perform clinical trials at earlier stages of treatment use, potentially improving recruitment. Greater awareness of factors affecting eligible patients’ participation in research is key to improving recruitment as well.

Improved integration of research-related and practice systems, specifically electronic medical records, could provide recruitment reminders and ease the integration of studies that require practice-billing systems. Electronic data systems with Internet ties would help to reduce data entry and analysis errors. Methodologically, as more data are collected and incorporated into cumulative analyses, subsequent trials may be modified. Using the population results, individuals may be able to undertake fewer crossovers and still obtain similar precision in their individual probability estimates. The cumulative capabilities could help achieve better precision as new questions arise.

Our work confirms the greater overall effectiveness of combined AMT+FL for improving IQ scores and provides support for further exploring the uses of practice-based N-of-1 trials for expanding effectiveness research. In bringing together community- and center-based physicians we showed that this approach was feasible, provided similar outcomes to a more standard trial design, and was well accepted by participants. Recruitment and enrollment posed challenges that we need to understand and address to improve feasibility and the generalizability of results of future studies. N-of-1 trials can be used in comparing a wide range of therapies, both standard and complementary, for many musculoskeletal and other chronic disorders. While providing individual-focused outcomes, the combined N-of-1 trial design offers promise for bridging “research in practice” and “practice to research.”

ACKNOWLEDGMENT
Special thanks to Dr. Roman Jaeschke for his most helpful input and oversight during this study. We thank Joanne Rhofiry and Nassima Lamriben for preparing the N-of-1 trial medication kits and Bonnie MacLeod for assistance with data management.

REFERENCES
Appendix: Bayesian analyses.

Bayesian methods enable the incorporation of prior information into inferences about the treatment effect. The Bayesian model employed a hierarchical structure with \( Y_{ij} \sim N(\alpha_i + \beta_i X_{ij}, \sigma_i^2) \), describing the distribution of scores \( Y_{ij} \) for the \( i \)th subject in the \( j \)th time period given treatment \( X_{ij} \) (\( X_{ij} = 1 \) if treated with AMT+FL; = 0 if AMT alone) and variances for each subject. The subject intercepts and slopes were modeled, respectively, as \( \alpha_i \sim N(\alpha_0, \tau_\alpha^2) \) and \( \beta_i \sim N(\beta_0, \tau_\beta^2) \). We also tested a fixed treatment model in which \( \beta_i \) was constant across individuals and considered a common variance Bayesian model (\( \sigma^2 = \sigma_i^2 \)).

Each model required specification of the prior probability distributions for: \( \alpha_0, \beta_0, \tau_\alpha^2, \tau_\beta^2, \sigma^2 \).

We separately used non-informative prior distributions as well as informative prior distributions based on results of the Goldenberg study (2). In that published crossover trial of 19 patients, scores on AMT ranged between 3 and 96 with mean 50 and standard deviation 23. The treatment differences ranged between -30 and 22 with mean -7.3 and standard deviation 17.5. We took the prior distribution for the mean intercept \( \alpha_0 \) (the mean score for AMT) to be normal with mean 50 and the standard deviation (SD) 4.6. Because \( \alpha_0 \) is a mean, this standard deviation is like a standard error and should not be confused with the standard deviation of the data. Likewise, the mean treatment difference \( \beta_0 \) between the AMT+FT and the AMT arms suggested a normal prior with mean 8 points and SD=3.9.

Information about the variance parameters \( \tau_\alpha^2 \) and \( \tau_\beta^2 \) was limited because only one measurement on each treatment was available in the prior data. Similarly, the few observations available for each patient in the prior study did not provide reliable data about the prior distributions for the within patient variances. Inverse gamma (IG) distributions are convenient computational choices for distributions of variances and we chose a fairly non-informative IG(1,8) prior for \( \tau_\alpha^2, \tau_\beta^2 \), and \( \sigma^2 \). This prior is centered near 12 with 25th and 75th percentiles of 6 and 28, but with a 99th percentile of 800 that allows for large variances as well. Note that a variance of 100 would correspond to a 40 point two standard deviation range about the treatment mean, so this should be a sufficiently vague prior given the bounds of the FIQ score between 0 and 100. The Bayesian models were fit using Markov chain Monte Carlo methods with the WinBUGS software.