

# N-of-1 Randomized Intervention Trials in Health Psychology: A Systematic Review and Methodology Critique

Jonathan A. Shaffer, PhD, MS<sup>1</sup> • Ian M. Kronish, MD, MPH<sup>2</sup> • Louise Falzon, PGDipInf<sup>2</sup> • Ying Kuen Cheung, PhD<sup>3</sup> • Karina W. Davidson, PhD<sup>2,4</sup>

Published online: 3 February 2018

© Society of Behavioral Medicine 2018. All rights reserved. For permissions, please e-mail: journals.permissions@oup.com.

## Abstract

**Background** Single-patient, multiple cross-over designs (N-of-1 or single-case randomized clinical trials) with systematic data collection on treatment effects may be useful for increasing the precision of treatments in health psychology.

**Purposes** To assess the quality of the methods and statistics, describe interventions and outcomes, and explore the heterogeneity of treatment effect of health psychology N-of-1 trials.

**Methods** We conducted a systematic review of N-of-1 trials from electronic database inception through June 1, 2015. Potentially relevant articles were identified by searching the biomedical electronic databases Ovid, MEDLINE, EMBASE, all six databases in the Cochrane Library, CINAHL, and PsycINFO, and conference proceedings, dissertations, ongoing studies, Open Grey, and the New York Academy's Grey Literature Report. Studies were included if they had health behavior or psychological outcomes and the order of interventions was

randomized. We abstracted study characteristics and analytic methods and used the Consolidated Standards of Reporting Trials extension for reporting N-of-1 trials as a quality checklist.

**Results** Fifty-four N-of-1 trial publications composed of 1,193 participants were included. Less than half of these (36%) reported adequate information to calculate the heterogeneity of treatment effect. Nearly all (90%) provided some quantitative information to determine the superior treatment; 79% used an a priori statistical cutoff, 12% used a graph, and 10% used a combination.

**Conclusions** N-of-1 randomized trials could be the next major advance in health psychology for precision therapeutics. However, they must be conducted with more methodologic and statistical rigor and must be transparently and fully reported.

**Keywords** N-of-1 trial • Within-subject design • Multiple cross-over design • Clinical decision making • Research methodology • Health psychology

✉ Karina W. Davidson  
kd2124@cumc.columbia.edu

<sup>1</sup> Department of Psychology, University of Colorado Denver, Denver, CO 80204

<sup>2</sup> Department of Medicine, Center for Behavioral Cardiovascular Health, Columbia University, 622 West 168th Street, New York, NY 10032

<sup>3</sup> Department of Biostatistics, Columbia University Mailman School of Public Health, Columbia University, New York, NY 10032

<sup>4</sup> Value Institute, New York Presbyterian, 5141 Broadway, New York, NY 10032

## Introduction

More than three decades ago, Guyatt and colleagues proposed an exciting, personalized method to solve the discrepancy between the results provided by large, conventional (between-subject) randomized clinical trials (RCTs) and the needs of individual patients or participants—the N-of-1 (single subject) study design [1, 2]. With the advent of comparative effectiveness research [3] and patient-centered outcomes research [4], there is a renewed interest in N-of-1 trials as an important research method for generating scientific evidence for what

is best for the single patient or participant [5]. Ironically, psychologists were performing this type of single-subject experimental design on psychological and behavioral phenomena for many decades before that [6, 7].

In an N-of-1 or patient-centered trial, the effect of one treatment is compared with one or more other treatments or a placebo condition, and the differences are calculated within-person. Thus, these designs are essentially multiple cross-over trials conducted on single persons [2]. Whereas conventional two-arm RCTs aim to estimate the average effect of an intervention in a population, an N-of-1 trial strives to identify the optimal treatment for a single patient or participant [8]. These trials may include observational and nonrandomized studies, such as ABAB designs, multiple baseline designs, and regression discontinuity designs. In the ideal version of this type of trial, a single person receives one or more reversible interventions and a credible control or comparator condition in a random order and with suitable consideration of a washout period. For example, a patient with chronic pain might compare acetaminophen versus ibuprofen, by taking each of them for 1 week at a time in a blinded

fashion, and then repeating the treatments two more times, while self rating pain and side effects on a daily basis over the course of the 6-week trial. At the end of the trial, the patient would review the unblinded results with their clinician and help choose which medication to continue to treat his or her pain. N-of-1 trials that involve randomization of the order of sequencing treatments are referred to here as N-of-1 RCTs.

N-of-1 RCTs are best suited to clinical problems for which there is substantial uncertainty or clinical equipoise about the optimal treatment. In addition, a number of characteristics ideally or necessarily need to be present to be able to conduct an N-of-1 RCT successfully (Table 1). Specifically, the treatment target in N-of-1 trials must exhibit some variation and must be measurable over time; thus, rapidly progressive conditions are not well suited to N-of-1 trials. The intervention to be tested must be able to be withdrawn, and its effects must be reversible—something that limits the behavior change techniques (BCTs) [9, 10] that can be tested within this design; the washout period between exposures to the intervention must be able to be estimated; and, at least

**Table 1** Considerations for Conducting N-of-1 Trials

	N-of-1 appropriate	N-of-1 not appropriate	Notes
State of knowledge	Clinical uncertainty or equipoise	Clear benefit of one intervention for the participant	
Nature of the problem	Chronic stable or slowly progressing, frequently recurring symptoms	Rapid disease, disorder, or problem progression	Time trends for symptoms should be considered in N-of-1 design (randomization vs. counterbalancing)
	Washout period between interventions is safe	Participant harm is possible if active intervention is discontinued	
Nature of interventions	Rapid efficacy	Slow efficacy onset	Masking of intervention assignment is ideal but not always necessary
	Minimal carryover across time	Substantial carryover across time (long washout)	
	Significant individual differences in intervention response are expected	Small individual differences in intervention response are expected	Consider different intervention efficacy onset periods in analyses
		Too complex/requires constant adjustment	
		No credible placebo	
Outcome assessment	Valid measures of outcome can be assessed multiple times	Primary outcome is assessed at a single time point	Repeated assessments within intervention condition periods are ideal
Statistical analytic plan	Appropriate statistical analyses (e.g., time-series analysis, Bayesian meta-analysis) can be conducted	Qualitative data are obtained	Statistician should be consulted prior to initiating N-of-1 trial
Willingness of stakeholders	Patient/participant, physician/investigator, pharmacist, and statistician willing to expend effort	One or more stakeholders not available/willing	All items must be conducted a priori, including obtaining institutional review board approval and patient/participant consent
Availability of financial resources	Measurement devices for outcome, cost of compounding drug, and multiple visits with physician can be procured	Resources not available	Most previous N-of-1 randomized clinical services have only existed where external grant support was available to cover these costs

in early comparative tests of an intervention, a credible placebo with similar expectancy effects or comparator intervention should ideally exist [11]. There should be a priori decisions about when and how outcomes will be measured; preferably, outcomes are measured objectively, and at minimum the assessors of the outcome are kept masked to the intervention assignment for that particular period. Ideally, participants and researchers are also kept masked to the condition that is offered during each of these periods; however, masking of those providing and receiving the intervention is often not possible for many trials that involve behavior or emotion change interventions. Finally, there should be a statistical approach, a power analysis, and a consideration of the data that would result in rejecting the null hypothesis articulated a priori. Of note, the sample size in an N-of-1 trial does not refer to the number of patients, but rather the number of measurements conducted by each patient as well as the number of periods in which each treatment is tested.

The primary advantage of the N-of-1 design is that it offers direct, objective evidence about the dose and/or efficacy of a particular intervention for a particular participant or patient [12]. These results can then be used to inform the treatment decision for that patient or participant. Engaging patients in data collection and interpretation can also enhance shared decision making during treatment selection, which may enhance patient satisfaction with care and may increase subsequent adherence with treatment. This differs from usual clinical practice in which information is extrapolated from a conventional randomized design involving other patients and not always generalizable to specific patients, and hence not useful to helping a specific participant and investigator make a decision about the best treatment or experimental option for her or him [13].

In a series of demonstration trials, N-of-1 trials have been confirmed to have the potential to lead to valuable changes in treatment, cessation of treatment, or confirmation of the original treatment [14–18]. For example, in one series of 71 completed N-of-1 trials for patients with chronic pain or osteoarthritis, 46 patients (65%) decided to change their pain medication as a result of the information from the trials, and of the 37 patients using a nonsteroidal anti-inflammatory drug or cyclooxygenase-2 inhibitor drug for pain management before their trials, 12 (32%) decided that the medication was not helping and stopped using it as a result of their trial results. Thus, N-of-1 trials can help individuals and their physicians empirically evaluate what is the best treatment for that patient.

N-of-1 trials might be especially well suited to helping patients find treatments that work to improve their health behaviors, ranging from low physical activity to medication adherence. First, there are likely to be substantial

differences in the underlying mechanisms for why different patients are nonadherent to recommended health behaviors. Given the challenges of identifying the specific mechanism in advance and then selecting the treatment, simply allowing a patient to test various promising treatments at the outset may be more helpful for finding the treatment that works for them. Second, the advent of mobile health devices such as actigraphers and electronic medication monitors can now facilitate the continuous and objective measurement of health behaviors. This has substantially reduced participant burden and increased feasibility and rigor of behavioral N-of-1 trials.

N-of-1 trials require the ability to collect sufficient outcomes data within a single patient to have enough power within-subject to compare differences in effect between treatments. Thus, N-of-1 trials are not well suited to prevention efforts relevant to health outcomes that infrequently occur, for example preventing migraine headaches that occur on, at most, a monthly basis. They are, however, well suited to comparing interventions for preventive health behaviors that can be continuously measured such as physical activity. Similarly, N-of-1 trials might be well suited to comparing interventions, at the individual level, that have the goal of maintaining health behaviors that are otherwise expected to rapidly decline over time. Interestingly, N-of-1 trials can also be useful for determining whether to discontinue a treatment. In our era of polypharmacy with many treatments having unclear benefit, N-of-1 trials that are designed to help dismantle a regimen may be a particularly promising use case.

A secondary benefit of N-of-1 trials is that they have the potential to provide different treatment estimates than those derived from between-subject trial designs or conventional parallel-arm RCTs. Evidence from conventional between-subject RCTs can conclude that the intervention tested was safe and effective on average but that particular intervention may have had an uneven mix of risks and benefits to individual patients, a problem quantified as heterogeneity of treatment effect (HTE) for those participants randomized to receive the intervention [19] and defined by Kravitz and colleagues as the “magnitude of the variation of individual treatment effects across a population” [20]. This broad definition also contains a more operational definition of HTE as the interaction of the treatment condition with individual participant characteristics. However, in conventional RCTs, there is limited power to detect such interactions, even in a meta-analysis combining the results from several studies [21]. There are exciting possibilities about the ways to calculate HTE from a series of N-of-1 trials pooled together that might yield a substantially different estimate than one calculated from a between-subject design or meta-analysis. One participant, for example, could have a medium-sized treatment effect that is positive, but if

that participant also had a medium-sized positive effect response to the placebo, there would be a different result from averaging this type of net treatment effect than what would be found in a between-subject trial design (although between-person trials do indeed have a similar problem in variation in control group and treatment as usual effects) [22]. Importantly, understanding when HTE is occurring with our health psychology interventions will help health psychology researchers, clinicians, and policymakers identify those interventions or treatments that might best be studied by N-of-1 trials, rather than continuing to be tested with a between-subject RCTs.

In contrast to conventional between-subject RCTs, the N-of-1 approach assumes that each person is sufficiently unique to warrant his/her own test of the usefulness of an intervention. However, one cannot determine whether this approach is generalizable for specific outcomes, interventions, or individuals unless multiple N-of-1 trials are conducted and their results pooled. Such a pooling would allow one to test whether the treatment effect is homogeneous or heterogeneous, and whether an N-of-1 RCT is necessary to identify which participants will predominantly benefit from this particular intervention or experimental condition. Thus, N-of-1 designs may allow for a more accurate translation of research findings and higher probability of increasing our understanding of who may benefit from many of our health behavior interventions.

All these N-of-1 trial considerations present a challenge for health psychology researchers, clinicians, and theorists. Medicine, particularly with pharmaceutical interventions, has data on most of these aspects before trials of any type are conducted. Health psychologists need to further their empirical research in this area to better specify which BCTs, theories, and other health psychology interventions can be appropriately tested in a N-of-1 RCT. To aid in the progress toward this goal, we conducted a systematic review to assess the quality of the methods and statistics used in N-of-1 trials, to describe the interventions and outcomes in N-of-1 trials for health behavior or psychological outcomes, and to explore the HTE in health psychology N-of-1 trials.

## Methods

The protocol for this systematic review was registered in PROSPERO before conducting the review and has also been previously published [23]. The reporting of this review conforms to Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) guidelines [24, 25].

Eligibility criteria for inclusion in this systematic review included the following [1]: all human populations

for whom N-of-1 trials with psychological or health behavior outcomes have been conducted [2]; all medical, psychological, and behavioral interventions for which N-of-1 trials have been conducted (i.e., no restrictions on interventions) [3]; placebo control or active treatment control; and [4] health behavior or psychological outcomes. Outcome categories included the following [5]: behavioral (e.g., number of steps) [6], self-reported subjective symptoms or severity (e.g., pain and dyspnea), and [7] psychological (e.g., depression). Interventions had to include randomization of treatment components (i.e., if an N-of-1 trial examining the relative efficacy of self-monitoring vs. goal setting for physical activity were conducted, the order in which these BCTs were administered had to be randomly assigned). Studies were excluded if they did not contain sufficient design detail to determine eligibility (i.e., they consisted primarily of methods and review without presentation of any data or results from an N-of-1 trial) and/or were not available in English.

## Search Strategy

Potentially relevant articles were identified by searching the biomedical electronic databases Ovid, MEDLINE, EMBASE, all six databases in the Cochrane Library, CINAHL, and PsycINFO from date of database inception through June 1, 2015. Conference proceedings, dissertations, ongoing studies, Open Grey, and the New York Academy's Grey Literature Report were also searched. All relevant subject headings and free-text terms were used to represent N-of-1 RCTs and psychological or behavioral interventions, and databases were searched from inception through the week of planned manuscript submission. Terms for MEDLINE included `n-of-1.tw OR ((individual or single) adj (patient$ or participant$ or subject$)).tw. OR ipd.tw AND exp Behavioral Medicine/ OR exp psychotherapy/ OR behavior$ adj (change or health or medicine or therap$)).tw OR psychotherap$.tw. OR psycholog$.tw`. These terms were adapted for the other databases. Ongoing studies were also sought through [Clinicaltrials.gov](http://Clinicaltrials.gov) and the World Health Organization International Clinical Trials Registry Platform. Additional records were identified by scanning the reference lists of relevant studies and reviews, by using the Related Articles feature in PubMed, and by using the Cited Reference Search in Scopus.

## Search Selection Process

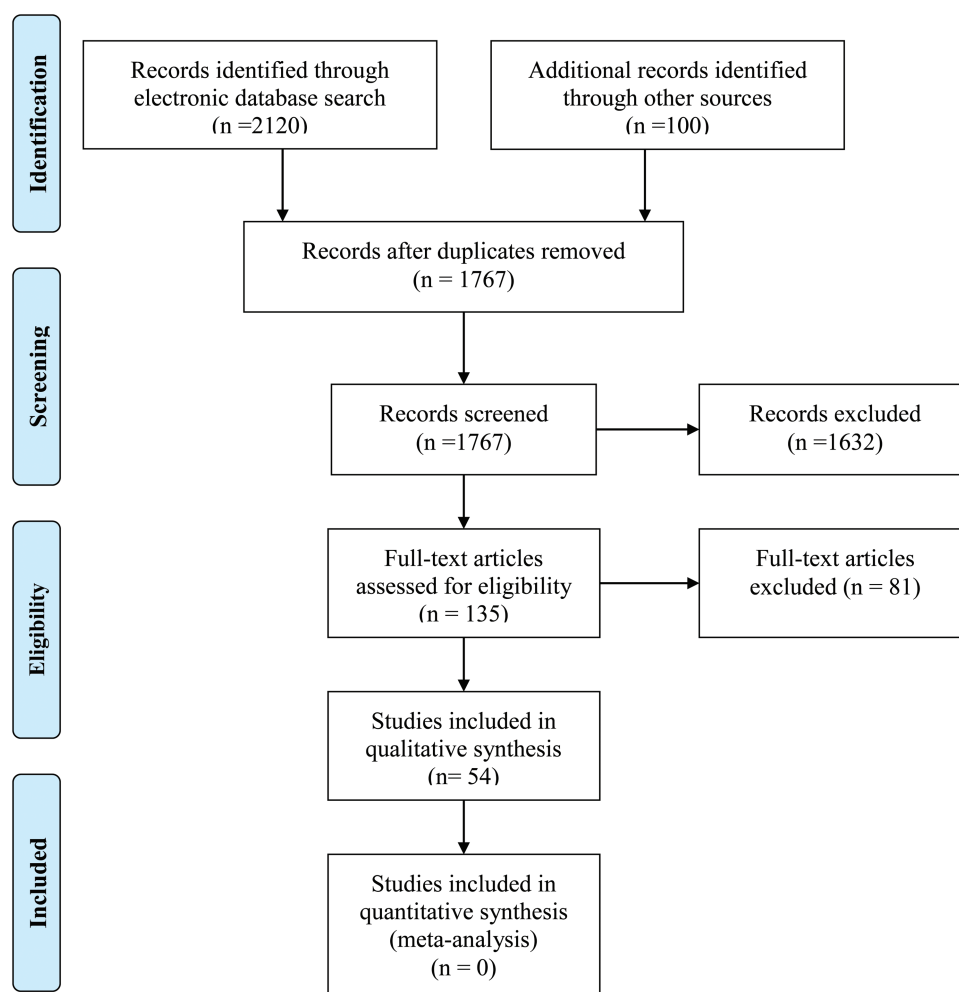
Two reviewers (J.A.S. and L.F.) independently screened titles and abstracts of all the retrieved bibliographic records. Full texts of potentially eligible records passing the title and abstract screening level were retrieved and examined independently by the two reviewers according

to the above-mentioned eligibility criteria. Disagreements at both screening levels (title/abstract and full text) were adjudicated by a third reviewer (K.W.D.). The study selection process with reasons for exclusions is shown in Fig. 1.

### Assessment of Study Quality

Assessment of study quality was performed by two reviewers (J.A.S. and L.F.) according to the Consolidated Standards of Reporting Trials (CONSORT) statement [24] and CONSORT Extension for N-of-1 trials (CENT) (Supplementary Table 1) [26]. The CENT checklist assesses aspects of a study's introduction, trial design characteristics, intervention, and outcome characteristics with sufficient detail to allow replication, allocation characteristics, statistical analytic methods, results, harms or unintended effects for the intervention, and discussion (trial limitations, generalizability, consideration of harms and benefits). Whereas the Template for Intervention Development

and Replication Checklist and Guide focuses exclusively on characteristics of the intervention featured in a study [27]. CENT aims to assess the entire study, including the intervention but also methods of randomization, statistical analytic methods, and discussion. Furthermore, CENT adds to the CONSORT diagram, addressing such topics as the rationale for using the N-of-1 approach, description and measurement properties of outcome assessment tools, and number of periods analyzed. Type of study end point (e.g., self-report rating scale, objectively observed behaviors) was recorded by two independent reviewers (J.A.S. and L.F.). As the goal was to focus on the quality but not the specifics of assessment tools, the precise outcomes measures were not extracted. Extraction data for 20% of studies were compared between reviewers to ensure accuracy of data extraction. Review of these outcomes was intended to provide clinicians and researchers with information that may be useful as they design their own N-of-1 trials by allowing them to identify clinical conditions and



**Fig. 1** Preferred Reporting Items for Systematic Reviews and Meta-Analyses flow diagram for N-of-1 health psychology trials



outcomes that may be particularly amenable to the N-of-1 method.

## Results

### Trial Description

We retrieved 1,767 unique articles that were potentially eligible for systematic review. Following title and abstract review and then full-text review, we selected 54 articles for inclusion in this systematic review. These 54 articles varied considerably in their content, addressing a variety of clinical conditions and interventions and using a variety of different methods. The health psychology–relevant conditions that were most commonly studied using N-of-1 trial designs included pain (15 articles), sleep and fatigue (13 articles), attention-deficit/hyperactivity disorder (8 articles), well-being and mood (6 articles), depressive symptoms (4 articles), emesis and nausea (4 articles), and cognitive function (4 articles) (Supplementary Table 2). By far the most common interventions tested were pharmacologic (41 articles). Behavioral interventions were tested in four articles: motivational interviewing for weight loss and physical activity [28], behavioral self-control [29], behavior modification [30], and goal setting and self-monitoring for physical activity [31]. Other types of interventions tested in N-of-1 trials included spinal cord stimulation versus transcutaneous electric nerve stimulation for pain [32], gastric fat infusions for sleep/fatigue and well-being [33], electrical stimulation for depressive symptoms [34], and varying food textures for increasing food intake [35].

### Quality of Methods Reported in These Trials

As indicated in Supplementary Table 1, study quality of the 54 articles selected for inclusion in this systematic review was generally poor. Less than half of all studies reported whether the design featured a run-in or washout period or specified concurrent conditions or interventions. Strikingly, only 15% of all studies discussed a power analysis or sample size determination or addressed statistical methods to account for carryover effects, intraclass correlation, or period effects. Similarly, few studies reported details regarding the implementation of study procedures, including who was masked, who generated the allocation sequence, who enrolled participants, and how the allocation sequence was concealed. In contrast, most studies described the scientific background under investigation (94%) and reported the trial design (98%) with planned number of periods (91%) and their duration (96%).

Few studies provided detailed information on the statistical methods that they used. In general, studies provided graphic or tabular presentation of results ( $n = 22$ );

used analysis of variance ( $n = 7$ ),  $t$ -tests ( $n = 23$ ), nonparametric tests ( $n = 10$ ), or mixed-effects regression models ( $n = 4$ ); and/or implemented hierarchical Bayesian models ( $n = 5$ ). Most studies (79%) used an a priori statistical cutoff, whereas the remainder used either a graph (12%) or a combination of graph and statistical test (10%) to determine treatment superiority. Less than half of these trials (40%) reported adequate information to calculate the HTE for a population or individual treatment effect (item 12b in Supplementary Table 2).

### Clinical or Scientific Usefulness Reported in These Trials

Only nine trials, which comprised 105 participants, reported on the usefulness of conducting these trials: 67% of these participants had subsequent treatment or scientific decisions consistent with the results of the trial (i.e., were prescribed the treatment that the N-of-1 trial identified as superior or non-inferior), 22% had decisions inconsistent with trial results, and 11% had ambiguous results.

### Quality Gaps in N-of-1 Designs With Psychological and Health Behavior Outcomes

The importance of a washout period separating active treatment periods has been a subject of contention [5]; although a washout period between treatment periods can be used to guard against carryover effects, patients and clinicians may think it is unethical or problematic to withhold active treatment during the washout period. Nonetheless, we found that some studies included a washout period. For instance, Huber et al. tested whether amitriptyline therapy in children aged 10–18 years with active polyarticular-course idiopathic arthritis results in significant improvement in pain compared with placebo [36]. They note that each 2-week course of treatment was separated from the next treatment by a 1-week washout period. Likewise, Louly et al., in their investigation of the effectiveness of tramadol compared with placebo for chronic cough, administered either 50 mg of tramadol twice daily or placebo, followed by a 2-day washout period, and then the alternate treatment [37]. Indeed, without a washout period between the two treatments, a residual effect of one active treatment might contribute to the outcome of a subsequent treatment. Indeed, carryover effects resulting from insufficient washout will often tend to reduce observed differences between treatments. Even if the benefits of a particular treatment wash out quickly, the risks of adverse treatment–related events can persist for some treatments (e.g., aspirin, which reduces pain in a matter of hours but increases risk of bleeding for up to 7 days) [38]. In these cases, the likelihood of detecting net benefit will depend, in part, on the order in which the treatments are administered,

and the N-of-1 trial results can be unduly biased by this ordering. N-of-1 trials can be designed to account for carryover effects by spacing out the intervals in which data pertaining to benefits and harms of treatments are assessed. However, this will lengthen the period needed for the full N-of-1 trial. N-of-1 trial designers must carefully balance what is needed to obtain rigorous, high-quality data with what is feasible for participants to complete. Similar problems arise when the onset of a new treatment is slow; in these cases, the N-of-1 trial design must provide sufficient time for a treatment to reach its maximal benefit before assessing treatment benefit. Statistical approaches are being developed that can help account for run-in and washout and effectively make N-of-1 trials more efficient [36].

Although nearly all the studies described in this review provided a general description of the trial design, none mentioned most of the design elements considered critical to the good conduct of an N-of-1 trial in the CENT quality checklist. For example, neither Anderson et al. [29] nor Gourlay et al. [39] identified their study as an N-of-1 trial in their trial; Coxeter et al. [40] and Eick et al. [41] failed to report the number of sequences performed; and Guyatt et al. did not report methods used to summarize data [2]. It is possible that these and other studies included this information and simply did not report it; however, we suspect that the authors' failure to report this information indicates a failure to consider these and other core N-of-1 trial features in their study design.

Few studies reported the method used to generate the allocation sequence, a core element of the CENT quality checklist for N-of-1 trials. In N-of-1 trials, randomization aims to “achieve balance in the assignment of treatments over time so that treatment effect estimates are unbiased by time-dependent confounders” [42]. Although randomization is not absolutely necessary for achieving balance, the studies we reviewed did not use alternative methods, such as the paired design (ABABABAB), the singly counterbalanced design (ABBAABBA), or the doubly counterbalanced design (ABBABAAB). The choice of these designs becomes important in the presence of systematic period effects. Few studies described the considerations of adjusting for period effects, although there are relatively straightforward methods that could safeguard against bias. For example, the singly counterbalanced design will guard against any linear time trend, whereas the paired design may lead to biased estimate of treatment effect for an individual. When there is a quadratic time trend, the doubly counterbalanced design may be used, whereas the singly counterbalanced design may not be adequate. Alternatively, post hoc statistical analysis can be easily applied to account for period effects that are additive to the treatment effects. Related study design elements, such as reporting the mechanism used

to implement the random allocation sequence, identifying steps taken to conceal the sequence until the interventions were assigned, and specifying the personnel responsible for generating the random allocation sequence, were also rarely featured in the articles that we reviewed.

In addition to sparse details regarding allocation of participants and randomization methods, few N-of-1 trial studies included detailed information on masking after assignment to interventions. For instance, Pelham et al. evaluated a behavioral intervention for eight hyperactive children for a period of 5 months [30]. Although it is impossible to mask the interventionists in this study, masking of the outcome assessors is possible. Nonetheless, outcome assessors in this study, who monitored on-task behavior in the classroom, were not masked. Although masking in N-of-1 trials may be less critical than in parallel group RCTs given that participants and clinicians are typically interested in the net benefit of treatment, including specific and nonspecific (i.e., placebo) effects, expert opinion tends to favor masking in N-of-1 trials whenever feasible [38]. Although masking is often not feasible in trials of behavioral interventions, most health psychology–relevant interventions identified by our systematic search involved pharmacotherapy, which makes masking through the use of a placebo more feasible.

Other core elements pertaining to the design of N-of-1 trials were infrequently reported in the 54 studies that we systematically reviewed. Specifically, less than 15% of studies provided dates defining recruitment and follow-up or information on whether any procedures were stopped early and/or whether the trial was stopped early. Few studies reported on the measurement properties of outcome assessments or on any changes to trial outcomes after the trial commenced. Although, as with each of the other design elements that studies failed to report, it is possible that outcome measures were indeed valid and reliable or that no changes to outcomes occurred after the trial commenced, this information should be reported to allow for optimally informed considerations of the validity of study findings.

### Quality Gaps in the Statistical Analysis of N-of-1 Trials With Psychological and Health Behavior Outcomes

Because most of the studies featured in our systematic review provided scant information regarding core N-of-1 design elements, so too did these studies fail to provide critical information pertaining to the statistical analysis of the trials. In particular, less than 20% of studies reported how sample size was determined (either number of participants or number of repetitions). Less than 10% of studies reported on any interim analyses and stopping guidelines, and less than 40% reported on methods of quantitative synthesis of individual trial data, including

adjusted analyses and how heterogeneity between participants was assessed. Furthermore, nearly all studies failed to provide information on statistical methods used to account for carryover effects, period effects, or intraclass correlation. Surprisingly, these studies also failed to report the results for each period for each primary and secondary outcome, the number of participants or periods analyzed, or the estimated effect size and precision for primary and secondary outcomes. This lack of information poses serious problems for the design and analysis of future N-of-1 trials because this information is needed to properly calculate sample size and number of repetitions of interventions.

## Discussion

This review found only 54 published studies, and most featured disparate outcomes and included rudimentary statistical analysis plans and design features. A total of 61% of studies targeted only three health psychology–relevant conditions: pain, sleep/fatigue, and attention-deficit/hyperactivity disorder. While a number of other conditions and problems (e.g., physical activity, quality of life, weight loss; Supplementary Table 1) were the target of at least some N-of-1 trials, N-of-1 trials have yet to be conducted for a number of other health psychology–relevant problems such as medication adherence, diet, anxiety, and panic attacks. A prior systematic review of N-of-1 trials in the medical literature found only 108 trials and identified similar deficiencies in the quality of the design and the reporting of those trials [43]. Although this previous review also sought to describe N-of-1 trial characteristics, examine treatment changes resulting from N-of-1 trial participation, and determine the adequacy of trial reporting, it focused exclusively on the medical literature, did not search several relevant databases (specifically, PsycINFO, the Cochrane databases, and CINAHL), and concluded in 2010.

The common deficiencies in design features that we identified included lack of consideration of a washout or run-in period and lack of consideration of concurrent participant health conditions or interventions. The majority of these N-of-1 studies were completed before the publication of the CONSORT extension for N-of-1 trials, and the deficiencies in N-of-1 study designs may have been driven by the lack of an established framework for designing and reporting high-quality N-of-1 trials. With respect to statistical aspects, less than 20% of studies featured a power analysis, and most used basic statistical techniques that do not take into account carryover effects, intraclass correlation, or period effects. This low use of a priori power analyses was likely driven by the fact that the parameters describing variability in outcome measures that are needed for N-of-1 analyses are typically only available

at the group level as opposed to the within-subject level. Thus, it is challenging to conduct power analyses until these data on heterogeneous treatment effects become more widely available.

## Recommendations to Improve the Quality of N-of-1 Designs With Psychological and Health Behavior Outcomes

We recommend that N-of-1 trials feature randomization or at least counterbalancing. These design elements protect against secular trends and other threats to validity that exist in our current crop of N-of-1 trials. If these design decisions are not possible (e.g., not acceptable to participants), then the study investigators should think carefully as to whether N-of-1 trials are an appropriate design for that particular scientific problem.

The use of valid and reliable outcome measures is also highly recommended because these measures provide one step toward thorough and systematic outcomes assessment. In choosing outcome measures, two issues are important to consider [1]: what data to collect and [2] how to collect them [38]. There is a tradeoff between choosing measures that are well validated but imprecisely related to participants' goals and measures that satisfy participants' priorities and are easy to complete.

Repetition of treatment exposure is also integral to the success of N-of-1 trials, and number of repetitions is analogous to sample size in conventional RCTs [38]. Although most studies included in our systematic review featured repetition of intervention administration, not all did. Indeed, some used the simplest N-of-1 trial design in which exposure to one treatment followed the other one time (AB or BA). One-time exposure to AB or BA offers limited protection against systematic error due to maturation and time-by-treatment interactions but no defense against random error [38]. These threats to the validity of N-of-1 trials are particularly important for many behavioral and psychological conditions that tend to wax and wane over time (e.g., depressive symptoms, dieting). Once again, if repetition is not feasible because of reasons such as burden/acceptability to participants conducting N-of-1 trials or inability to reverse effect of interventions, then N-of-1 trials may not be appropriate.

## Recommendations to Improve the Statistical Analytic Methods for N-of-1 Trials With Psychological and Health Behavior Outcomes

Our finding that most studies used less advanced statistical tests to analyze N-of-1 trial data is consistent with a previous review by Gabler et al., who found that approximately half of trials reported using a *t*-test or other simple statistical criterion, whereas another



half reported using a visual/graphic comparison alone [43]. We assessed some studies that used more sophisticated Bayesian methods to analyze trial data [5, 44], and we recommend that these alternate approaches be explored further and used in future N-of-1 trials. Furthermore, statistical methods should take into account possible intervention carryover effects if washout and run-in periods are not integrated into the study design.

### **Facilitating a Quantitative Meta-Analysis of N-of-1 Trial Designs With Psychological and Health Behavior Outcomes**

We were unable to conduct a quantitative meta-analysis of the 54 studies of N-of-1 trials with psychological or health behavior outcomes given the disparate outcomes and analytic methods and the minimally reported information on core design elements. Several pieces of information are needed to conduct a quantitative meta-analysis in the future. This information includes data for each primary and secondary outcomes *for all periods* of the N-of-1 trial; results of inferential statistical tests rather than simply qualitative descriptions or visual presentation of results; information on study participants, including concurrent conditions and treatments; and information on core design elements, such as those addressed above (presence or absence of a washout and run-in period, information pertaining to allocation sequence, etc). Ideally, investigators should use the CENT checklist to guide the design, analysis, interpretation, and reporting of their N-of-1 trials. Moreover, for the field to progress, it may be useful to develop a repository of N-of-1 trial designs with clearly delineated N-of-1 trial designs and outcome measures that investigators can reuse such that the results of individual trials can be easily pooled together.

### **How the Increase in mHealth Will Allow Health Psychologists to be Leaders in This Design**

Excitingly, with the advent of smartphones and other mHealth devices, there are now novel tools that can be used to collect data pertaining to outcomes that are relevant to many behavioral and psychological conditions. Accelerometers, for example, can be used to capture data pertaining to physical activity. Smartphone diaries can be used to ecologically capture outcome data pertaining to psychological symptoms. There are even innovative approaches to using mHealth sensors to capture outcome data pertaining to emotional states [45]. Such approaches make it easier for participants to collect data than the paper and pencil surveys that were used in the past. These tools also make it possible to directly input data into N-of-1 trial electronic platforms that can

integrate the data and produce real-time evidence comparing treatments.

### **A Primer on Conducting N-of-1 Trials for Psychological and Health Behavior Outcomes**

First and foremost, it is essential to determine whether the N-of-1 method is applicable to the clinical question of interest (Table 1). Indications include substantial clinical uncertainty, chronic or frequently recurring symptomatic conditions, and treatments with rapid onset and minimal carryover. Contraindications include rapidly progressive condition, treatment with slow onset or prolonged carryover, and participant or investigator insufficiently interested in reducing therapeutic uncertainty to justify effort.

Second, after establishing the applicability of the N-of-1 method to the clinical question, one must select trial duration, treatment period length, and a sequencing scheme. In so doing, one might consider that longer trial duration offers greater precision but may be difficult or tedious to complete with the potential for dropout during the trial. Furthermore, treatment period length should be adjusted to fit the therapeutic half-life (of drug treatments) or treatment onset and duration (of nondrug treatments). Finally, although simple randomization optimizes masking, counterbalanced sequencing is generally a more reliable guarantor of validity given the limited number of repetitions that are possible within the sequence of treatments without overburdening participants conducting the N-of-1 trials.

Third, one should implement a washout period if indicated. Washout is sometimes more acceptable and indicated if treatment duration of action is short relative to treatment period and contraindicated if participants could be harmed by cessation of active treatment. For instances in which withholding treatment is unethical, one can use an analytic approach that down weights, disregards, or does not collect outcomes at the beginning of a treatment period [46]. Other modeling approaches are possible. Researchers should also consider data trends in N-of-1 trials (e.g., increases in step counts) attributable to increased intervention exposure and learning over time. Masking is feasible for pharmacologic interventions but challenging for most nondrug treatments (behavioral, lifestyle). Adequate masking allows investigators to distinguish between specific and nonspecific treatment effects, although in some circumstances, this distinction might not matter to participants and investigators.

Fourth, one should select suitable outcome domains and measures, although one should also consider that patient-centeredness should not be trumped by psychometric considerations [38]. Of note, outcome domains and measure must also be validated and easily repeatable in order to ensure N-of-1 trials are feasible from a

participant burden and validity of results point of view. Finally, measures that have been used in prior N-of-1 trials may be preferred if pooling of results to understand population level HTE is desired.

Fifth, one must analyze and present data to support clinical decision making by participants and investigators. A reasonable approach is to select one or two primary outcome measures that best represent benefits and harms. For clinician-prescribed treatments, data should be presented using a variety of both statistical and graphical methods to fully explicate the data in a participant-friendly manner that facilitates shared decision making with their clinicians while allowing for the calculation of sample sizes or number of repetitions in future N-of-1 studies.

Sixth, ethical considerations for N-of-1 trials are also important. There should not be a clear clinical preference for or against one of the treatments based on known health outcomes (equipose) [38]. In addition, one must consider whether informed consent and institutional review board (IRB) approval are required. In practice, clinicians often test treatments, measure outcomes, and determine whether to continue or change treatment, in what has been referred to as a “trial and error” approach. Clinicians do not obtain IRB approval for every clinical decision. If clinicians were to incorporate the more scientific N-of-1 method into this process, purely for the purpose of guiding treatment decisions for that individual patients (and not for publication), then many would argue that it would be ethical to pursue the N-of-1 approach to patient care without IRB involvement. We envision a future in which N-of-1 trials become more routinely incorporated into care pathways, without the burdensome requirements of IRB approval. Of course, if N-of-1 trials are designed to aggregate and disseminate the results, then IRB involvement may be necessary.

### Limitations of the Current Review

Notwithstanding the timely nature of this comprehensive review, some limitations deserve mention. First, our search strategy did not include questionnaires or scales, which commonly are specifically developed and validated for self-report outcomes for outcome evaluations for specific conditions (e.g., depression scales). Second, we have not provided detailed information on how to statistically analyze N-of-1 trials (e.g., by discussing which intraclass correlations are necessary and how to conduct time-series analyses). We believe this discussion is beyond the scope of the current paper, and we refer the reader to other literature (e.g., Duan et al. [38]). Third, our initial search strategy did not include the term single case or single patient. We corrected for this oversight

by conducting an additional search strategy with these terms and found that none of the articles recovered were eligible for inclusion in this review.

### Conclusions

Increasing recognition, based on statistical, methodological, and empirical grounds, points to serious problems with between-subject designs to test health psychology-relevant interventions and outcomes, particularly when HTE is large. These issues are particularly pernicious for the investigation of areas of interest to health psychology because universal laws are unlikely to operate for many of the behaviors or processes we wish to change. Furthermore, even if exemplar universal behaviors exist, they are more likely at the level of reflexes, operating before exposure to environmental, learning, and other idiosyncratic influences. Aggregating well-designed N-of-1 trials may help us more definitively answer the question as to whether certain behavioral treatments are broadly useful or whether ideographic approaches are essential to achieving large treatment effects in health psychology-relevant problems. Efforts to compare the N-of-1 approach to usual practice will help us learn which areas of health psychology can best avail themselves of this personalized approach.

### Supplementary Material

Supplementary material is available at *Annals of Behavioral Medicine* online.

**Acknowledgments** This work was supported by the National Heart, Lung, and Blood Institute under Grants K23 HL112850, K23 HL098359, R01 HL123368, P01 HL088117, P01 HL047540, and K24 084034 (J.A. Shaffer, K.W. Davidson, I.M. Kronish, L. Falzon); Agency for Healthcare Research and Quality under Grant R01 HS024262 (I.M. Kronish); National Library of Medicine under Grant R01 LM012836 (K.W. Davidson); Patient-Centered Outcomes Research Institute under Grant ME-1403-12304 (J.A. Shaffer, K.W. Davidson, L. Falzon); and National Institute of Neurological Disorders and Stroke Grant R01 NS072127 (Y.K. Cheung).

### Compliance with Ethical Standards

**Conflict of Interest:** The authors have no conflicts of interest to declare.

**Authors' Contributions:** Dr. Karina Davidson was responsible for the conceptualization and design of this study, and contributed the development of the manuscript and its editing. Dr. Jonathan Shaffer contributed to the development of and revisions to the manuscript. Dr. Ian Kronish contributed to the development of and revisions to the manuscript. Ms. Louise Falzon conducted the search strategy and assessment of study quality with Dr. Shaffer. Dr. Ying Keun Cheung contributed to the statistical content of the manuscript.

**Ethical Approval:** No new human subjects research was conducted as part of this manuscript, so ethical approval was not obtained.

**Informed Consent:** For this type of study formal consent is not required.

## References

- Guyatt G, Sackett D, Adachi J, et al. A clinician's guide for conducting randomized trials in individual patients. *CMAJ*. 1988; 139(6): 497–503.
- Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S. Determining optimal therapy—randomized trials in individual patients. *N Engl J Med*. 1986; 314(14): 889–892.
- Sox HC, Greenfield S. Comparative effectiveness research: a report from the Institute of Medicine. *Ann Intern Med*. 2009; 151(3): 203–205.
- Selby JV, Beal AC, Frank L. The Patient-Centered Outcomes Research Institute (PCORI) national priorities for research and initial research agenda. *JAMA*. 2012; 307(15): 1583–1584.
- Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J Clin Epidemiol*. 2013; 66 (suppl 8): S21–S28.
- Ebbinghaus M. *Memory*. New York, NY: Columbia Teacher's College; 1913.
- Angell J, Pierce A. Experimental research upon the phenomena of attention. *Am J Psychol*. 1891; 4(4): 528–541.
- Sedgwick P. “N of 1” trials. *BMJ*. 2012; 344: e844.
- Michie S, Richardson M, Johnston M, et al. The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Ann Behav Med*. 2013; 46(1): 81–95.
- Abraham C, Michie S. A taxonomy of behavior change techniques used in interventions. *Health Psychol*. 2008; 27(3): 379–387.
- Davidson KW, Peacock J, Kronish IM, Edmondson D. Personalizing behavioral interventions through single-patient (N-of-1) trials. *Soc Personal Psychol Compass*. 2014; 8(8): 408–421.
- Barlow D, Hersen M. *Single Case Experimental Designs: Strategies for Studying Behavior Change*. 2nd ed. New York, NY: Pergamon Press Inc.; 1984.
- Duan N, Kravitz RL, Schmid CH. Single-patient (n-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. *J Clin Epidemiol*. 2013; 66 (suppl 8): S21–S28.
- Guyatt G, Jaeschke R, McGinn T. N-of-1 randomized controlled trials: study design. In Guyatt G, Rennie D, Meade MO, Cook DJ, eds. *Users' Guides to the Medical Literature*. 2nd ed. Alphen aan den Rijn, Netherlands: McGraw Hill Medical; 2008: 179–192.
- Guyatt GH, Keller JL, Jaeschke R, Rosenbloom D, Adachi JD, Newhouse MT. The N-of-1 randomized controlled trial: clinical usefulness. Our three-year experience. *Ann Intern Med*. 1990; 112(4): 293–299.
- Nikles CJ, Yelland M, Glasziou PP, Del Mar C. Do individualized medication effectiveness tests (n-of-1 trials) change clinical decisions about which drugs to use for osteoarthritis and chronic pain? *Am J Ther*. 2005; 12(1): 92–97.
- Larson EB. N-of-1 trials: a new future? *J Gen Intern Med*. 2010; 25(9): 891–892.
- Joy TR, Monjed A, Zou GY, Hegele RA, McDonald CG, Mahon JL. N-of-1 (single-patient) trials for statin-related myalgia. *Ann Intern Med*. 2014; 160(5): 301–310.
- Greenfield S, Kravitz R, Duan N, Kaplan SH. Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. *Am J Med*. 2007; 120(4 suppl 1): S3–S9.
- Kravitz RL, Duan N, Braslow J. Evidence-based medicine, heterogeneity of treatment effects, and the trouble with averages. *Milbank Q*. 2004; 82(4): 661–687.
- Altman DG, Bland JM. Interaction revisited: the difference between two estimates. *BMJ*. 2003; 326(7382): 219.
- de Bruin M, Viechtbauer W, Hoppers HJ, Schaalma HP, Kok G. Standard care quality determines treatment outcomes in control groups of HAART-adherence intervention studies: implications for the interpretation and comparison of intervention effects. *Health Psychol*. 2009; 28(6): 668–674.
- Shaffer J, Falzon L, Cheung YK, Davidson KW. *N-of-1 Randomized Trials in the Health Psychology Literature: A Systematic Review and Methodology Primer PROSPERO [CRD42015017853]*. York, UK: University of York Centre for Reviews and Dissemination; 2015.
- Moher D, Liberati A, Tetzlaff J, Altman DG; PRISMA Group. Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. *BMJ*. 2009; 339: b2535.
- Moher D, Shamseer L, Clarke M, et al.; PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. *Syst Rev*. 2015; 4(1): 1.
- Vohra S, Shamseer L, Sampson M, et al.; CENT Group. CONSORT extension for reporting N-of-1 trials (CENT) 2015 Statement. *BMJ*. 2015; 350: h1738.
- Hoffmann TC, Glasziou PP, Boutron I, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. *BMJ*. 2014; 348: g1687.
- Vail-Gandolfo N. The effects of a motivational program on glycemic control, physical activity and weight loss in individuals with type 2 diabetes mellitus. *Dissertation Abstracts Int, Sec B: Sciences Engineering*. 2009; 70(1-B): 703.
- Anderson EE, Clement PW, Oettinger L Jr. Methylphenidate compared with behavioral self-control in attention deficit disorder: preliminary report. *J Dev Behav Pediatr*. 1981; 2(4): 137–141.
- Pelham WE Jr, Carlson C, Sams SE, Vallano G, Dixon MJ, Hoza B. Separate and combined effects of methylphenidate and behavior modification on boys with attention deficit-hyperactivity disorder in the classroom. *J Consult Clin Psychol*. 1993; 61(3): 506–515.
- Snihotta FF, Presseau J, Hobbs N, Araújo-Soares V. Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trials. *Health Psychol*. 2012; 31(6): 733–737.
- Cepeda MS, Acevedo JC, Alvarez H, Miranda N, Cortes C, Carr DB. An N-of-1 trial as an aid to decision-making prior to implanting a permanent spinal cord stimulator. *Pain Med*. 2008; 9(2): 235–239.
- Wells AS, Read NW, Macdonald IA. Effects of carbohydrate and lipid on resting energy expenditure, heart rate, sleepiness, and mood. *Physiol Behav*. 1998; 63(4): 621–628.
- Tellez-Zenteno JF, McLachlan RS, Parrent A, Kubu CS, Wiebe S. Hippocampal electrical stimulation in mesial temporal lobe epilepsy. *Neurology*. 2006; 66(10): 1490–1494.
- DeMatteo C, Law M, Goldsmith C. The effect of food textures on intake by mouth and the recovery of oral motor

- function in the child with a severe brain injury. *Phys Occup Ther Pediatr*. 2002; 22(3–4): 51–71.
36. Huber AM, Tomlinson GA, Koren G, Feldman BM. Amitriptyline to relieve pain in juvenile idiopathic arthritis: a pilot study using Bayesian metaanalysis of multiple N-of-1 clinical trials. *J Rheumatol*. 2007; 34(5): 1125–1132.
  37. Louly PG, Medeiros-Souza P, Santos-Neto L. N-of-1 double-blind, randomized controlled trial of tramadol to treat chronic cough. *Clin Ther*. 2009; 31(5): 1007–1013.
  38. Kravitz RL, Duan N, eds. *Design and Implementation of N-of-1 trials: A User's Guide*. Rockville, MD: Agency for Healthcare Research and Quality; 2014.
  39. Gourlay S, Forbes A, Marriner T, Kutin J, McNeil J. A placebo-controlled study of three clonidine doses for smoking cessation. *Clin Pharmacol Ther*. 1994; 55(1): 64–69.
  40. Coxeter PD, Schluter PJ, Eastwood HL, Nikles CJ, Glasziou PP. Valerian does not appear to reduce symptoms for patients with chronic insomnia in general practice using a series of randomised n-of-1 trials. *Complement Ther Med*. 2003; 11(4): 215–222.
  41. Eick TJ, Kofoed L. An unusual indication for a single-subject clinical trial. *J Nerv Ment Dis*. 1994; 182(10): 587–590.
  42. Kravitz RL, Duan N, Vohra S, Li J; the DECIDE Methods Center N-of-1 Guidance Panel. Introduction to N-of-1 trials: indications and barriers. In: Kravitz RL, Duan N, eds. *Design and Implementation of N-of-1 Trials: A User's Guide*. Rockville, MD: Agency for Healthcare Research and Quality; 2014: 1–11.
  43. Gabler NB, Duan N, Vohra S, Kravitz RL. N-of-1 trials in the medical literature: a systematic review. *Med Care*. 2011; 49(8): 761–768.
  44. Schluter PJ, Ware RS. Single patient (n-of-1) trials with binary treatment preference. *Stat Med*. 2005; 24(17): 2625–2636.
  45. Saeb S, Zhang M, Karr CJ, et al. Mobile phone sensor correlates of depressive symptom severity in daily-life behavior: an exploratory study. *J Med Internet Res*. 2015; 17(7): e175.
  46. Hogben L, Sim M. The self-controlled and self-recorded clinical trial for low-grade morbidity. *Br J Prev Soc Med*. 1953; 7(4): 163–179.