original article

A starter kit for undertaking n-of-1 trials

Felix Naughton
University of Cambridge
Derek Johnston
University of Aberdeen

Preface

The aim of this article is to provide readers who have not yet undertaken n-of-1 or within-subject experimental studies with a general overview of the methodology from a health psychology perspective and to provide some tools to give readers the opportunity to give it a go themselves.

Introduction

The population based randomised controlled trial (RCT) has dominated intervention evaluation for many decades. However, one important downside of this general design is that it provides only an estimate of the *average* effect of an intervention for a given population. Although subgroup analyses within RCT samples are potentially informative, they fall short at being able to explain whether an intervention works for individual participants or small discrete groups of participants. There are also limitations with using group or population experiments to test psychological theory. Identifying relationships between theoretical constructs across individuals does not inform us on whether these relationships hold within individuals (Johnston & Johnston, 2013), which is arguably a valuable, perhaps essential, feature of any theory of behaviour. N-of-1 studies can generate evidence for the impact of an intervention or relationship between theory-derived constructs for specific individuals and identify inter-individual differences in these observations. Why is this valuable? For several reasons.

N-of-1 studies, because they use regular and numerous measurements within individuals, can provide good evidence for directions of causality. For example, whether exposure to an intervention precedes and explains changes in self-efficacy, which in turn precedes and explains changes in behaviour (potentially via intention or goal). N-of-1 RCTs also provide an opportunity to test discrete components of interventions, such as Behaviour Change Techniques (BCTs) (Michie et al., 2013), on behavioural determinants and behaviour between and within individuals (Craig et al., 2008) without the large samples required in population studies. This includes factorial n-of-1 randomised controlled trials which vary treatments on multiple occasions within individuals to identify their impact on short-term changes in behaviour (Sniehotta, Presseau, Hobbs, & Araujo-Soares, 2012). Importantly, with the smartphone becoming ubiquitous, data collection for these studies can be undertaken relatively easily and efficiently. This includes Ecological Momentary Assessment, an approach for collecting within-individual data in a person's naturalistic environment in real time (Shiffman, Stone, & Hufford, 2008).

What is an n-of-1 RCT?

An n-of-1 RCT is a crossover experiment conducted with a single participant who acts as their own control. Multiple n-of-1 RCTs can be aggregated statistically in order to explore between-participant as well as within-participant effects (see discussion section). N-of-1 RCTs usually provide repeated and randomly allocated periods of treatment to participants with sufficient frequency to minimise any chance of confounding influences on the

outcome. Furthermore, n-of-1 RCTs are often undertaken 'double-blind' where both the participant and researcher collecting data are blinded from treatment allocation, although this is frequently not possible in psychology studies.

According to the American Medical Association's Evidence Based Medicine Working Group, n-of-1 trials are regarded as the gold standard for generating evidence for individual treatment decisions, over and above systematic reviews of randomised controlled trials, and can provide definitive evidence of treatment effectiveness in individuals (Guyatt et al., 2000). However, only certain types of intervention and behavioural or health outcomes of interest in health psychology and related fields are appropriate for n-of-1 RCTs.

What types of interventions or outcomes are n-of-1 RCTs suitable for?

For interventions, a key issue in assessing whether n-of-1 RCTs are suitable is whether the intervention is likely to generate substantial carryover effects. If an intervention aims to change an individual's beliefs to bring about some change in their behaviour, through using persuasion say, any belief changes could last well beyond a crossover to a different intervention. In this scenario it can be difficult to determine whether any changes in behaviour after the persuasion intervention had ended was due to any subsequent intervention or due to the carryover effects of the original persuasion intervention. Therefore, interventions expected to produce only short-lasting effects on the outcome of interest, such as planning, goal setting, contingent reinforcement or rewards, self-monitoring and feedback interventions, as Sniehotta et al. (2012) suggest, are most suitable for n-of-1 RCTs as their carryover effects can be minimised. Similarly, investigating the impact of drug interventions including treatment efficacy, withdrawal or side-effects is particularly suitable. The blinding of participants and researchers is usually straightforward with drug related trials and carryover effects can be managed, providing appropriate 'wash-out' periods are factored in. When interventions have very substantial and/or enduring effects, other n-of-1 designs can be used, including multiple baseline designs where different behaviours are targeted sequentially or stepped wedge designs in which different participants have pre-intervention periods of different durations.

In terms of outcomes, those easily measured over short periods of time which are good predictors of longer term behaviour or clinical outcomes, are most suitable for n-of-1 RCTs e.g. abstinence from smoking. When investigating outcomes relating to specific health conditions, the stability of that condition can affect the ease to which changes in outcomes can be attributed to specific interventions. So stable conditions are most suitable for n-of-1 RCTs.

N-of-1 RCT case study

This next section of the article will describe a case study of an n-of-1 RCT undertaken to test a specific hypothesis about the experience of caffeine withdrawal for one individual. After the case study section, a description will be provided of how the analysis was undertaken and output interpreted with links to the actual data collected and analysis syntax to enable readers to undertake their own analyses for training purposes.

Hypothesis

PD [pseudonym] will experience caffeine withdrawal when her once-daily cup of caffeinated coffee is replaced with decaffeinated coffee.

Design

A single participant (n-of-1) double-blind randomised controlled trial of caffeinated versus decaffeinated coffee. Treatments were randomly allocated to twelve randomly selected treatment period blocks of 3 or 4 days (see allocation sequence in figure 1). Simple urn randomisation without

Treatment period	Period 1	Period 2	Period 3	Period 4	Period 5	Period 6	Period 7	Period 8	Period 9	Period 10	Period 11	Period 12
Allocation	A	B	B	A	B	B	B	A	A	B	A	A
Day	1 2 3 4	5 6 7	8 9 10	11 12 13 14	15 16 17	18 19 20	21 22 23	24 25 26	27 28 29 30	31 32 33	34 35 36	37 38 39 40

Figure 1: Allocation sequence for the caffeine case study

replacement was used to generate the allocation sequence using WinBUGS software (Lunn, Thomas, Best, & Spiegelhalter, 2000), undertaken by the statistician (DL). The researcher (FN) and the participant (PD) were blinded to allocation but only the participant was blinded to the treatment blocks.

Procedure

A single-blind manipulation check prior to the study demonstrated that the participant was unable to distinguish between caffeinated and decaffeinated coffee with added milk. During the 40-day study period PD was provided with the allocated treatment (caffeinated or decaffeinated coffee with milk) once a day in the mid-morning as per usual consumption and was discouraged from consuming other food or drink which contained caffeine. Nominated colleagues and friends, who were blinded from allocation, made the coffee at work and home respectively for the participant. The coffee was stored in identical tins labelled A and B. Nominated colleagues/friends were informed every morning by SMS text message about PD's treatment allocation (A or B) for that day using a free automated text message programme for Android (SMS Scheduler). The participant completed a study questionnaire at approximately 4pm every day during the study period either on their mobile phone or a pc.

Measures

The primary outcome measure was the mean score on the Caffeine Withdrawal Symptom Scale (CWSQ) (Juliano, Huntley, Harrell, & Westerman, 2012). Secondary outcomes were three subscales of the CWSQ, mood disturbance, decreased sociability and headache, selected as symptoms the participant felt she had experienced prior to the study shortly after abstinence from caffeine.

The participant was also asked to indicate on the daily questionnaire whether she believed she had consumed a caffeinated or decaffeinated coffee earlier that day, using a 5-point rating scale (from 'sure it was caffeinated' [1] to 'sure it was decaffeinated' [5]), whether they experienced any treatment violations (i.e. didn't drink a study coffee that day) and whether they had consumed any other food or drink containing caffeine that day. The participant could also add comments about their day which were considered relevant to the study using a free text field. Additional measures included perceived stress, sleep quality, alcohol consumption and minutes of vigorous physical activity.

Statistical analyses

Firstly, the CWSQ scale and subscale scores across the 40-day study period were plotted using SPSS. Secondly, we investigated whether these outcomes exhibited autocorrelation in SPSS. We then investigated whether allocation predicted scores on the CWSQ scale and subscales, when taking into account autocorrelation, using McKnight et al.'s double bootstrap method (McKnight, McKean, & Huitema, 2000). Finally, logistic regression was undertaken to assess whether the participant predicted, above chance, which treatment she was allocated to each day and linear regression was undertaken to assess whether allocation continued to predict CWSQ scores when the participant's assessment of which treatment she was receiving was taken into account.

Results

An essential first step in an n-of-1 study is to plot the data [*to create plot see A1 in the next section*].

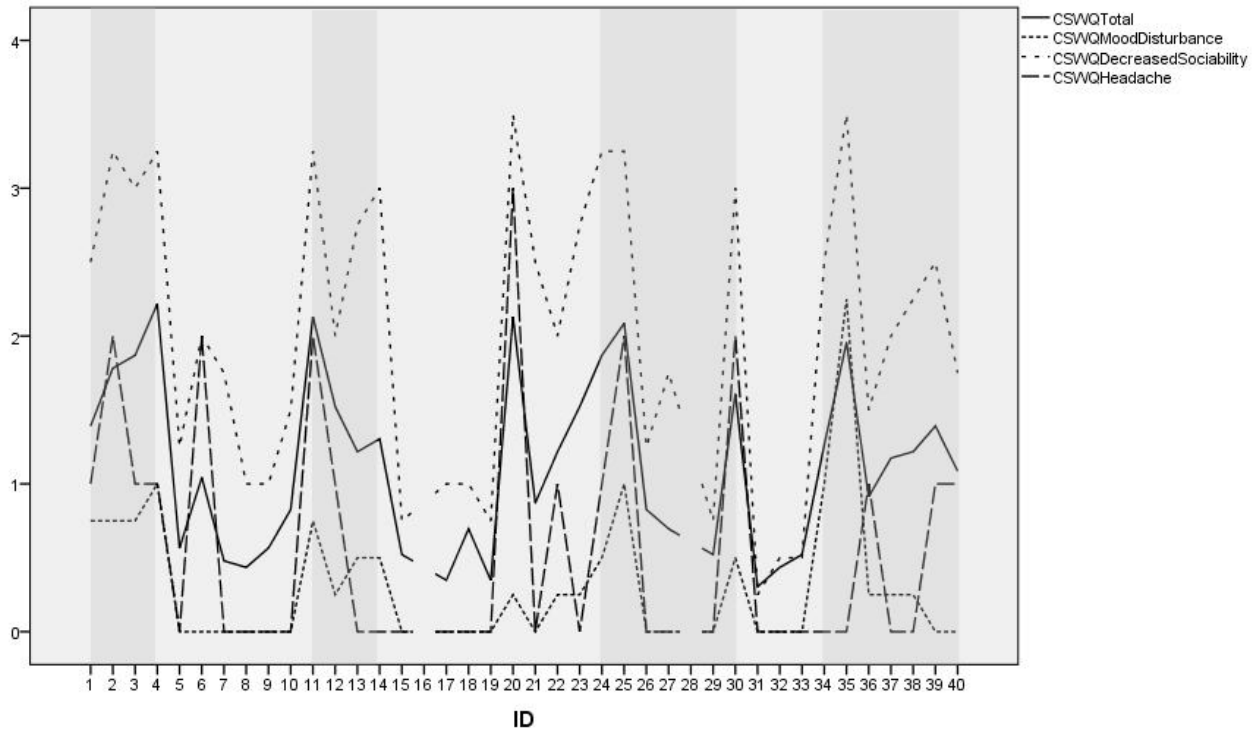

Figure 2: Plot showing the daily scores on the CWSQ and selected subscales over the 40-day study period (with missing data). Dark grey sections represent treatment periods where the participant received decaffeinated coffee

Figure 2 contains a plot showing the daily scores on the CWSQ and selected subscales over the 40-day study period. Overlayed in grey are the decaffeinated coffee treatment blocks when withdrawal symptoms, as measured by the CWSQ, are hypothesised to be higher. There were three treatment violations (days 14, 20 and 35), where the participant did not have a study coffee, and two days with missing data (days 16 and 28). The average value for the treatment block was substituted for the missing data. It is very likely that successive readings in an n-of-1 study will be correlated (autocorrelated, see glossary) a feature that can lead to inaccurate estimates of statistical significance. The CWSQ and subscales did not demonstrate significant autocorrelation [*A2*], although the mood disturbance subscale autocorrelation approached significance (figure 3). However the intervention could mask an underlying autocorrelation. This is allowed for in the analysis we used.

In a form of regression analyses designed for n-of-1 studies which we describe below [*A3*], treatment allocation predicted scores on the CWSQ (unstandardised beta estimate -0.74, $p < 0.001$), and the mood and decreased sociability subscales. As indicated in figure 2, there were two days (6 and 20) where scores on the CWSQ scale and subscales spiked, demonstrating increased withdrawal symptoms, despite being during a caffeine treatment period. When examining the additional information collected on the study questionnaire, the participant had indicated that these two days followed excessive alcohol consumption episodes the day before ("hungover") and on day 20 the study treatment was missed out, which was meant to be a caffeinated coffee. These appear to explain these unexpected spikes in withdrawal symptoms, taking into account the general similarity between symptoms of alcohol hangovers and caffeine withdrawal (Finnigan, Hammersley, & Cooper, 1998). The participant performed better than chance at predicting which treatment she had been allocated to that day (beta

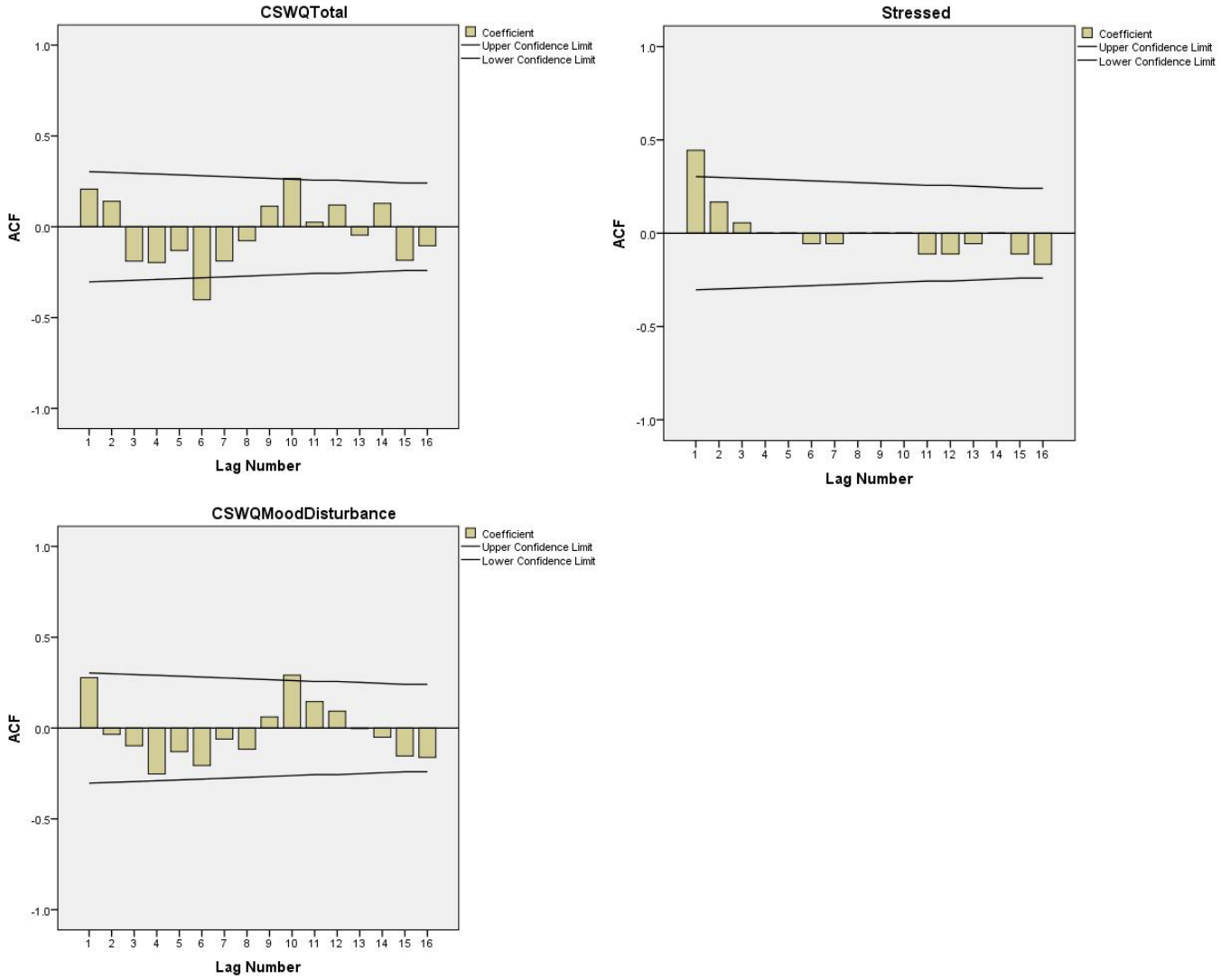

Figure 3: Autocorrelation charts for CWSQ total, mood disturbance CWSQ subscale and perceived stress

-1.26, p = 0.002), although allocation remained a significant predictor of scores on the CWSQ when her prediction was taken into account. The participant described guessing which treatment she had received based on how she felt later on in the day after consuming the treatment coffee in the morning rather than basing it on the taste, smell or appearance of the coffee.

Conclusion

The trial generated evidence that PD experiences caffeine withdrawal when caffeinated coffee, drunk on a one-a-day basis, is replaced with decaffeinated coffee.

Undertaking the analysis and interpreting the output

With the participant's permission, we have made the data we collected for this study freely available to enable others to use it and replicate our analyses for training purposes (and potentially do further exploratory analyses). We have provided SPSS syntax

for creating the scale and subscale variables from the raw data and to carry out the SPSS-related analyses described in the case study. We have also formatted the data into ASCII so it can be used with McKnight et al.'s double bootstrapping web-tool.

Files made available at https://osf.io/zp93r/files/ for use as part of this starter kit paper include raw data in CSV format (file 1), raw data in Excel format (file 2), transformed and coded data in SPSS v.21 format (file 3), SPSS syntax (file 4), CWSQ scale summary score data formatted for the McKnight software (file 5) and a guide on time series analysis of n-of-1 data using the prewhitening approach in SPSS (file 6).

Main analyses undertaken in caffeine case study

A1 - To plot the data in SPSS (as in figure 2), go to Analyse -> Forecasting -> Sequence charts and select your variable of interest and enter the time/date variable into the time axis label field.

A2 - To assess autocorrelations in SPSS, (including a graph as in figure 3), go to Analyse -> Forecasting -> Autocorrelations and select your variable of interest (you can leave the default options as they are).

A3 - To undertake McKnight's double bootstrap method go to the website (http://www.stat.wmich.edu/slab/Software/Timeseries.html)[1] . The data is entered as an ASCII file (.txt) with the data for each measurement period being on a separate row and the final data point in each row being the dependent (outcome) variable, all the other variables being assumed to be independent (predictor variables). See the above link for the caffeine study data in the required format (file 5). We find it best to cut and paste the dataset into the space provided in the web tool. Unlike SPSS and other major packages the constant (intercept) has to be specified. This done by entering 1 and it is conventional to make it the first variable. The other variables specify the experimental conditions and any other covariates that you may wish to use. We find that many people initially find it helpful and reassuring to first specify and run the regression model (with no allowance for autocorrelation) in whatever statistical package they normally use[2]. The software requires one to specify the degree of autocorrelation one wishes to allow for. First order (see autocorrelation in glossary) is the default and is a good starting point. The output from the double bootstrapping software provides estimates of the unstandardised beta weights, associated standard errors and tests of significance. The output also contains information on variances and covariances that can be ignored at least initially and estimates of the autocorrelation that was established and allowed for in the analyses. See figure 4 for an edited example of output from the web-tool.

Discussion

The remainder of the article provides some general rules of thumb about designing and analysing n-of-1 RCTs.

Aggregating n-of-1 trials

There are several ways to aggregate data from multiple n-of-1 trials, including meta-analysis and multi-level modelling (MLM). We favour MLM. Aggregating n-of-1 RCTs using these approaches enables the assessment of the overall or average effect of an intervention for a group of participants.

[1] At the time of publishing 7 November 2014 the server holding this software was not available. We understand that a new operating system is being installed, when completed the software will be available again. We also understand that an R version of the software is near completion. Despite these current uncertainties we have chosen to present our analyses using this software since it appears the best available option for dealing with small n-of-1 data sets. If large data sets are available then the ARIMA modelling procedures available in most statistical packages can be used.

[2] It is possible to make some allowance for autocorrelation by prewhitening the outcome variable and using the prewhitened variable as the outcome in a regression analysis. Instructions for doing this in SPSS (produced by Karen Schroder and Diane Dixon) can be found in additional material file 6 and syntax to analyse the caffeine study data using this approach is in the SPSS syntax file 4.

Time Series Results

Parameter Estimates and Test that parameter is zero

Parameter	Estimate	t-ratio	p-value	Authors' comments
Beta 1 (constant)	1738	6.92	<0.00001	Constant significant (minor importance)
Beta 2 (time)	-157	-1.72	0.0938	Slight effect of time
Beta 3 (intervention)	-742	-3.79	0.0006	Significant effect of the intervention

Variance Covariance Matrix of Parameter Estimates

Beta 1	Beta 2	Beta 3
0.6295	*-0.0190*	*-0.2883*
	0.0008	*0.0046*
		0.3843

Diagonal indicates variance of beta
Covariances in *italics*

Bootstrap Estimates and CI's of AR Parameters

Bootstrap Residual MSE = **.262397**		
Parameter	**Estimate**	**95% CI**
AR 1	.102	-.312, .517

No significant first order autocorrelation

Variance Covariance Matrix of AR Estimates

0.41833

Figure 4: Edited and annotated output from McKnight et al.'s double bootstrap method

With sufficient n-of-1 RCTs, it is possible to compare the effect of interventions on individuals with different characteristics. The use of MLM in n-of-1 studies is well described by Shadish, Kyse, and Rindskopf (2013).

Determining the number of data points, and number and length of treatment blocks

A key question asked with n-of-1 RCTs is how many data points are required. Ideally this should be based on what would provide sufficient power to detect the predicted or clinically significant difference between conditions. This would be dependent upon the nature of the outcome and intervention (Lillie et al., 2011). Sniehotta et al. (2012) applied Cohen's rule of thumb of having at least 30 participants per condition to provide 80% power. So for their n-of-1 RCTs this was translated into 30 data points per study condition. Ultimately, the more conditions/treatment periods there are, the greater the reduction of any potential confounding effects of other factors or behaviours on the outcome of interest. In terms of the length of treatment blocks, this very much depends on the length of time which one would expect an intervention to affect the outcome and cease affecting the outcome after it is removed. For the case study above, caffeine withdrawal is expected to start after 12 to 24 hours after caffeine abstinence and peak after 1-2 days. Withdrawal ceases rapidly once caffeine consumption resumes. Therefore treatment blocks of 3 or 4 days were deemed sufficient to capture caffeine withdrawal symptoms. However, interventions with long 'wash-out' periods or which take a significant amount of time to influence the outcome will require longer treatment periods and in some cases would not be suitable for n-of-1 RCTs. Practical considerations will often determine the number of observations possible in each replication of a treatment as well as the number of replications.

Testing for carryover effects

N-of-1 RCTs are most suitable for interventions with minimal carryover effects. But how do you know

if an intervention has a carryover effect? The first question to ask is whether an intervention is aiming or expected to produce anything more than a short-term effect on the individual. One rule of thumb suggested by Sniehotta et al. (2012) for assessing carryover effects after undertaking an n-of-1 RCT(s) is to see if there is an overall time trend i.e. does the outcome increase or decrease from the beginning of the study to the end. They also suggest that, for studies with very short treatment blocks e.g. one day, the existence of autocorrelation of the outcome could also be a weak indicator of carryover effects. Examination of the plot of the data is very helpful in detecting carryover effects.

Randomisation

In general, it is advisable to randomise the sequence of treatment blocks (Lillie et al., 2011). However, one downside of using simple randomisation is the risk that all treatment blocks end up clustered together. Therefore, where possible, some form of block randomisation is advisable to address this issue unless there are a large number of replications. In the above caffeine study example, we used a slightly different approach - simple urn randomisation without replacement. This is where exactly six treatment periods for each treatment were placed into a virtual urn and then selected at random in turn. Each time a treatment is 'pulled out' of the urn and selected for allocation to a treatment block, the probability of selecting the alternative treatment rises. This approach is considered to increase the unpredictability of allocation compared to permuted-block designs (Schulz & Grimes, 2002), although it does not entirely eliminate the risk of all treatment blocks of the same treatment ending up together in a row.

Conclusion

While N-of-1 RCTs have in the past predominantly been used to inform individual patient treatment, they offer utility for intervention development and evaluation in health psychology. There is evidence that their use to evaluate health interventions is increasing, partly driven by the increased practicality for both researchers and participants of collecting data via mobile digital devices. There still remains much debate as to how best to design n-of-1 studies. However, this can be overcome with greater use and exploration of this methodology. With the increased focus in health psychology on the specific 'active' components of interventions, n-of-1 trials may have an important role to play in this exciting new chapter of behavioural science.

Useful resources

Kravitz, R. L., Duan N. (Eds), & the DEcIDE Methods Center N-of-1 Guidance Panel (Duan, N., Eslick I., Gabler, N.B., Kaplan, H. C., Kravitz, R. L., Larson, E. B., Pace, W. D., Schmid, C. H., Sim, I., Vohra, S.) (2014). Design and Implementation of N-of-1 Trials: A User's Guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, MD: Agency for Healthcare Research and Quality. Retrieved from http://www.effectivehealthcare.ahrq.gov/search-for-guides-reviews-and-reports/?pageaction=displayproduct&productid=1844

The European Health Psychology Society (EHPS) n-of-1 Special Interest Group (open to any researchers who want to engage with others interested in and using n-of-1 designs): currently located at http://ehps.net/synergy/?q=node/135

Glossary

Autocorrelation: The association between sequential data points within the same variable. If data is collected daily (as with the above caffeine withdrawal study), the autocorrelation will examine the correlation between a variable at T0 and T-24hrs (lag 1) and then between T0 and T-48hrs (lag 2) and so on always going back in time. For a 1st order autocorrelative (or autoregressive) relationship, there will be an association at lag 1 but very little else at

further lags after that first association is taken into account.

Crossover period: The transition where one intervention is stopped and another intervention or non-intervention phase starts.

Crossover effect: When the effect of an intervention lasts beyond the point at which that intervention is withdrawn.

Washout period: A period to allow any crossover effects to cease before a separate intervention is provided.

Acknowledgements

A special thanks goes to PD for participating and being happy for her data to be made available, to Dave Lunn for advising on randomisation and creating the randomisation sequence and to Caren Schroder and Diane Dixon for allowing us to use bits from their SPSS guide on time series analysis and make it publically available.

References

Craig, P., Dieppe, P., Macintyre, S., Michie, S., Nazareth, I., & Petticrew, M. (2008). Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ, 337*, a1655. doi: 10.1136/bmj.a1655

Finnigan, F., Hammersley, R., & Cooper, T. (1998). An examination of next-day hangover effects after a 100 mg/100 ml dose of alcohol in heavy social drinkers. *Addiction, 93*(12), 1829-1838. doi:10.1046/j.1360-0443.1998.931218298.x

Guyatt, G. H., Haynes, R. B., Jaeschke, R. Z., Cook, D. J., Green, L., Naylor, C. D., ... Richardson, S. (2000). Users' Guides to the Medical Literature: XXV. Evidence-based medicine: principles for applying the Users' Guides to patient care. Evidence-Based Medicine Working Group. *JAMA, 284*(10), 1290-1296. doi:10.1001/jama.284.10.1290

Johnston, D. W., & Johnston, M. (2013). Useful theories should apply to individuals. *British Journal of Health Psychology, 18*(3), 469-473. doi:10.1111/bjhp.12049

Juliano, L. M., Huntley, E. D., Harrell, P. T., & Westerman, A. T. (2012). Development of the caffeine withdrawal symptom questionnaire: caffeine withdrawal symptoms cluster into 7 factors. *Drug and Alcohol Dependence, 124*(3), 229-234. doi:10.1016/j.drugalcdep.2012.01.009

Lillie, E. O., Patay, B., Diamant, J., Issell, B., Topol, E. J., & Schork, N. J. (2011). The n-of-1 clinical trial: the ultimate strategy for individualizing medicine? *Personalised Medicine, 8*(2), 161-173. doi: 10.2217/pme.11.7

Lunn, D. J., Thomas, A., Best, N., & Spiegelhalter, D. (2000). WinBUGS - a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing, 10*(4), 325-337. doi:10.1023/A:1008929526011

McKnight, S. D., McKean, J. W., & Huitema, B. E. (2000). A double bootstrap method to analyze linear models with autoregressive error terms. *Psychological Methods, 5*(1), 87-101. doi:10.1037/1082-989X.5.1.87

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... Wood, C. E. (2013). The Behavior Change Technique Taxonomy (v1) of 93 Hierarchically Clustered Techniques: Building an International Consensus for the Reporting of Behavior Change Interventions. *Annals of Behavioral Medicine, 46*(1), 81-95. doi:10.1007/s12160-013-9486-6

Schulz, K. F., & Grimes, D. A. (2002). Generation of allocation sequences in randomised trials: chance, not choice. *The Lancet, 359*(9305), 515-519. doi:10.1016/S0140-6736(02)07683-3

Shadish, W. R., Kyse, E. N., & Rindskopf, D. M. (2013). Analyzing data from single-case designs using multilevel models: new applications and some agenda items for future research. *Psychological*

Methods, 18(3), 385-405. doi:10.1037/a0032964
Shiffman, S., Stone, A. A., & Hufford, M. R. (2008). Ecological momentary assessment. *Annual Review of Clinical Psychology, 4*, 1-32. doi:10.1146/annurev.clinpsy.3.022806.091415
Sniehotta, F. F., Presseau, J., Hobbs, N., & Araujo-Soares, V. (2012). Testing self-regulation interventions to increase walking using factorial randomized N-of-1 trials. *Health Psychology, 31*(6), 733-737. doi:10.1037/a0027337

Felix Naughton
University of Cambridge, UK
fmen2@medschl.cam.ac.uk

Derek Johnston
University of Aberdeen, UK
d.johnston@abdn.ac.uk