BJPsych Advances (2024), vol. 30, 322–325 doi: 10.1192/bja.2023.65

RESEARCH METHODS

Should we compromise on *n*-of-1 data analyses?

Prathiba Batley

Prathiba Batley, PhD, is Director of Statistical Innovation at Daiichi Sankyo, Louisville, Kentucky, USA. Her research interests lie in single-case designs, Bayesian statistics and latent variable modelling. She has over 60 publications in top peer-reviewed outlets and over 90 presentations.

Correspondence Prathiba Batley. Email: pnbatley@gmail.com

First received 3 Jul 2023
Final revision 11 Nov 2023
Accepted 26 Nov 2023

SUMMARY

Despite their increasing popularity, *n*-of-1 designs employ data analyses that might not be as complete and powerful as they could be. Borrowing from existing advances in educational and psychological research, this article presents a few techniques and references for rigorous data analytic techniques in *n*-of-1 research.

KEYWORDS

n-of-1 designs; single-case experimental designs; data analysis; statistical methodology; Bayesian.

Although most medical research is concerned with differences in group means or trends in a group, modelling, analysing and understanding inter-individual differences are as important as intra-individual differences. For instance, when studying the person-specific effects of individualised cognitive therapy for depression in women with metastatic cancer (Lévesque 2004), pharmacological treatment for depression (Kronish 2018), stimulants for attention-deficit hyperactivity disorder in children (Mordijck 2018) or treatments for schizophrenia (Marwick 2018) the clinician is more interested in learning whether the personalised treatment has a positive impact on the patient. In fact, studying group effects may say nothing about how an individual progresses.

Consider, for example, a study that evaluates the efficacy of cognitive therapy for depression among women with metastatic cancer. Conducting a randomised control trial (RCT) in this case would involve recruiting women with metastatic cancer and randomly assigning them to experimental and control groups. There are several problems with this approach. First, recruiting a sample that is large enough to provide sufficient power in an RCT could be prohibitive when it comes to women with metastatic cancer. Second, there could be intra-individual variations that might not be accounted for. This could have a detrimental effect on validity. Third and most important, what works for one patient might not work for another. In such a case, finding a group mean would be misleading and would be associated with a large standard error. Additionally, the arithmetic average is not the most representative statistic in multimodal distributions. In fact, Schork (2015) gives examples of 10 most used drugs in America whose efficacy ranges from helping 1 in 4 individuals to 1 in 25! Finally, a snapshot measure of depression would provide much poorer information about the progress of depression than studying the individual across time. An article by Zuidersma et al (2020) provides more detailed information on why single-case experimental designs (SCEDs) or *n*-of-1 designs are appropriate and most suited for psychiatry research. The purpose of this article is not to show why or how they are important in psychiatric research, but to provide a primer for related data analytic standards that could be of value to that research.

In contrast to RCTs, using an individual as his or her own control (using baseline observations) and comparing against this baseline, observing data patterns across time, and controlling for baseline and treatment condition provide for much richer and more nuanced data. This is exactly what *n*-of-1 designs (as they are called in medicine), which are special cases of single-case experimental designs (SCEDs, as they are called in educational and psychological research), are designed to address. Although there are significant advances made in setting the standards and analytical tools for SCEDs, the current conduct of *n*-of-1 designs leaves much to be desired. We can attribute the lack of *n*-of-1 studies or their proper conduct to the myths about *n*-of-1 designs, such as the study involving only one participant, insufficient power, unavailability of statistically sophisticated analytical approaches and lack of knowledge transfer across subject domains. By the lack of knowledge transfer across subject domans, I mean that there are several tools and standards available not just in *n*-of-1 designs but also in SCEDs from which the medical research community can benefit. An example is the What Works Clearinghouse standards (ies.ed.gov/ncee/wwc/Handbooks), which discuss the design and analytical standards for SCEDs that can help provide the highest level of evidence of treatment effect. In the present article I focus on how to analyse SCED data using the

WWC standards as a guide. I explain some available statistical methods that mental health researchers can use to safeguard against various threats to validity in *n*-of-1 designs.

N-of-1 designs

N-of-1 designs are especially important in experimental designs where randomisation or collecting large data-sets are inappropriate or impossible, such as in mental health research, rare diseases or comorbid conditions. Contrary to popular belief, *n*-of-1 designs could include multiple participants and the question of power in SCEDs with small samples has been addressed in recent research (e.g. Hedges 2023). As mentioned above, methods currently used to analyse these data in medical research leave much to be desired. The parent design for *n*-of-1 designs, the SCED, has gained prominence in educational and psychological research and has a rigorous set of standards set out by the What Works Clearinghouse (ies.ed.gov/ncee/wwc/Handbooks). There have been several methodological developments in SCED data analysis in educational and psychological research. Although there are recommended standards for *n*-of-1 designs (Porcino 2020), these do not include how data should be analysed to address issues of validity. A recent systematic literature review (Natesan Batley 2023a) shows that of the 115 *n*-of-1 medical research articles published in the past 10 years, only 4 met the criteria of the WWC standards and only one study reported an appropriate effect size.

WWC standards: analytical criteria

The WWC standards require that the following analytical criteria be met to declare a SCED as providing strong evidence of treatment effect: (a) documenting consistency of level, trend and variability within a phase; (b) documenting immediacy of the effect; (c) reporting an appropriate effect size; and (d) ruling out external factors and anomalies. Level refers to the phase mean, trend refers to the slope of the phase and immediacy refers to how immediately the treatment takes effect when introduced or the data return to baseline following removal of the treatment. Additionally, there must be at least three demonstrations of the treatment effect to provide strong evidence. To achieve this, we examine level, trend, variability, immediacy, overlap/effect size and consistency of data patterns across phases. In this article we will consider only these six pieces of evidence that can be estimated statistically.

Data analytic considerations

SCED and, by extension, *n*-of-1 design data are often autocorrelated and have small sample size. This deadly combination renders most commonly used parametric analyses inappropriate. In SCEDs it has long been recognised that estimating and interpreting autocorrelations is important, and ignoring autocorrelations leads to incorrect estimates and inflated type I error rates (Huitema 2000). Therefore, the model that is used must include and account for autocorrelations. Researchers have shown that the combination of autocorrelated and small sample data can be effectively handled using the Bayesian framework (Natesan Batley 2023b).

Immediacy, consistency and variability

Researchers could use the Bayesian unknown change-point (BUCP) model (Natesan 2017) to measure immediacy. In BUCP modelling, the change point, that is, the time point associated with the introduction or the removal of the treatment, is assumed to be unknown. The algorithm for an interrupted time-series design estimates the 'unknown' change point along with the autocorrelation and the phase means and, if necessary, the phase slopes. The strength of the evidence of the change-point estimate (that is, the accuracy and its credible interval width) can be used to determine immediacy. This approach allows the data to speak for themselves. It is possible that in some *n*-of-1 designs, particularly in psychiatry, immediate effect might not be possible. In fact, most psychiatric treatments take at least a few weeks to reach efficacy. In those cases of latency, the change-point estimate can be used to show when the treatment started taking effect following its introduction or completely stopped taking effect following its removal. Credible intervals of Bayesian estimates of phase means and regression coefficients of phases speak to the consistency and the variability of the data between and within phases. No statistically rigorous standards exist on determining how much consistency and variability are permitted. This is an avenue for future research.

Level and trend

An analysis of small sample *n*-of-1 data that simultaneously estimates level, trend and autocorrelation is often underpowered (Natesan Batley 2021). Therefore, researchers must choose whether there is reason to believe that the data show a trend and then decide on an appropriate model. This could mean that the researcher chooses at most two of the three statistics (level, trend, autocorrelation) to estimate. The above-mentioned BUCP models can be used for these

analyses. For instance, in a psychiatry research study that is measuring the impact of an antidepressant on post-traumatic stress disorder (PTSD), the means of the phases and either the slopes of the phases (if there is a reason to believe in a trend across time) or the autocorrelation could be estimated. The differences between the means (levels) across phases and the differences between the slopes across phases (trends) could be reported.

Effect sizes

The crux of any experimental design is to estimate a design-comparable and appropriate effect size. The effect size must be appropriate in the sense that it should consider the scale of the data and the autocorrelation, and have an appropriate small sample correction. The effect size must be design-comparable in the sense that it must be comparable across studies to facilitate meta-analytic work. However, a systematic review found that 99.1% of the studies used incorrect effect sizes, such as the mean difference, Cohen's *d* and *R* (Natesan Batley 2023a). There are non-overlap indices in SCEDs, but these are not design-comparable and they suffer from other statistical drawbacks. Two solutions (effect size computations) exist so far for ABAB and multiple baseline designs, proposed by Hedges et al (2012, 2013). Consider the previous example of a psychiatry research study that is measuring the efficacy of an antidepressant on PTSD. A researcher could use the phase means, the autocorrelations and the intraclass correlations to compute these effect sizes. These effect sizes are particularly suited for psychiatry and *n*-of-1 studies because it is not uncommon for such studies to include at least three patients. This facilitates the use of multilevel models to compute intraclass correlations. The caveat, of course, is that computing these effect sizes requires some statistical rigor, but this can be relatively easily overcome using free software such as R and packages such as lme4. However, these effect sizes are appropriate only for continuous data – commonly used ordinal scale data require an extension of the odds ratio effect size for SCEDs.

Obviously, this brief article cannot discuss in detail power considerations in *n*-of-1 designs. Hedges et al (2023) showed that in ABAB designs more phase reversals with fewer observations per phase have more power than fewer phase reversals with more observations per phase. That is, if the researcher can obtain only 36 observations, an ABABAB design with 6 observations per phase has higher power than an ABAB design with 9 observations per phase. Also see Natesan Batley (2023c) on this. This is another emerging area of research that is essential in designing *n*-of-1 studies.

Conclusion

All *n*-of-1 designs must ensure highest quality of evidence of treatment effect that is uncompromised; but the ultimate purpose of *n*-of-1 designs is to be able to be meta-synthesised across studies. For both purposes, appropriate analytical techniques are necessary. Currently, it seems that *n*-of-1 researchers are not sufficiently focused on using an appropriate index to quantify treatment efficacy. This article has presented some techniques that have shown rigor in analysing SCEDs and, by extension, *n*-of-1 designs. However, more methodological research in *n*-of-1 designs is necessary to expand the applicability and robust conduct of this promising experimental design.

Acknowledgement

I thank Dr Nicholas John Batley for his consultation on the medical science aspects of this article.

Funding

This work was funded by the Institute of Education Sciences (grant number R305D220052).

Declaration of interest

None.

References

Hedges LV, Pustejovsky JE, Shadish WR (2012) A standardized mean difference effect size for single case designs. *Research Synthesis Methods*, **3**: 224–39.

Hedges LV, Pustejovsky JE, Shadish WR (2013) A standardized mean difference effect size for multiple baseline designs across individuals. *Research Synthesis Methods*, **4**: 324–41.

Hedges LV, Shadish WR, Natesan Batley P (2023) Power analysis for single-case designs: computations for $(AB)^k$ designs. *Behavioral Research Methods*, **55**, 3494–503.

Huitema BE, McKean JW (2000) Design specification issues in time-series intervention models. *Educational and Psychological Measurement*, **60**: 38–58.

Kronish IM, Hampsey M, Falzon L, et al (2018) Personalized (*N*-of-1) trials for depression: a systematic review. *Clinical Psychopharmacology*, **38**: 218–25.

Lévesque M, Savard J, Simard S, et al (2004) Efficacy of cognitive therapy for depression among women with metastatic cancer: a single-case experimental study. *Journal of Behavioral Therapy and Experimental Psychiatry*, **35**: 287–305.

Marwick KFM, Stevenson AJ, Davies C, et al (2018) Application of *n*-of-1 treatment trials in schizophrenia: systematic review. *British Journal of Psychiatry*, **213**: 398–403.

Mordijck E, Danckaerts M, Onghena P (2018) [N-of-1 trials in child and adolescent psychiatry: a closer look at stimulants]. *Tijdschrift Voor Psychiatrie*, **60**: 315–25.

Natesan P, Hedges LV (2017) Bayesian unknown change-point models to investigate immediacy in single case designs. *Psychological Methods*, **22** (4): 743.

Natesan Batley P, Hedges LV (2021) Accurate model vs. accurate estimates: a study of Bayesian single-case experimental designs. *Behavior Research Methods*, **53**: 1782–98.

Natesan Batley P, McClure EB, Brewer B, et al (2023a) Evidence and reporting standards in N-of-1 medical studies: a systematic review. *Translational Psychiatry*, **13**(1): 263.

Natesan Batley P (2023b) Bayesian analysis of single case experimental design count data in trauma research: a tutorial. *Psychological Trauma*, **15**: 829–37.

Natesan Batley P, Thamaran M, Hedges LV (2023c) AB^{k}PowerCalculator: an app to compute power for balanced (AB)^k single case experimental designs. *Multivariate Behavioral Research*. [Epub ahead of print] 17 Oct. Available from: https://doi.org/10.1080/00273171.2023.2261229.

Porcino AJ, Shamseer L, Chan AW, et al (2020) SPIRIT extension and elaboration for n-of-1 trials: SPENT 2019 checklist. *BMJ*, **368**: m122.

Schork NJ (2015) Time for one-person trials. *Nature*, **520**: 609–11.

Zuidersma M, Riese H, Snippe E, et al (2020) Single-subject research in psychiatry: facts and fictions. *Frontiers in Psychiatry*, **11**: 539777.