Jane Nikles · Geoffrey Mitchell *Editors*

The Essential Guide to N-of-1 Trials in Health

Springer

The Essential Guide to N-of-1 Trials in Health

Jane Nikles • Geoffrey Mitchell
Editors

The Essential Guide to N-of-1 Trials in Health

Springer

Editors

Jane Nikles
School of Medicine
The University of Queensland
Ipswich, QLD, Australia

Geoffrey Mitchell
School of Medicine
The University of Queensland
Ipswich, QLD, Australia

ISBN 978-94-017-7199-3 ISBN 978-94-017-7200-6 (eBook)
DOI 10.1007/978-94-017-7200-6

Library of Congress Control Number: 2015952660

Springer Dordrecht Heidelberg New York London

Printed on acid-free paper

Springer Science+Business Media B.V. Dordrecht is part of Springer Science+Business Media (www.springer.com)

Preface

The fundamental problem facing the clinical research enterprise is this: what clinicians (and patients) want to know is not what clinical trials are equipped to say. Randomized controlled trials (RCTs) are all about the average patient; they yield average treatment effects. Doctors and patients want individual treatment effects: how a given patient will respond to Treatment A versus Treatment B. No amount of statistical subterfuge can make standard-issue, parallel group RCTs reveal precisely the results we want. There is one method, however, that under certain conditions can reliably identify the best treatment for an individual. That method is the N-of-1 trial.

N-of-1 trials are crossover experiments conducted with a single patient. They are applicable principally to treatments used for symptomatic, chronic, nonfatal conditions. By systematically observing a patient's response to two or more treatments, we can determine which of the treatments are likely to work best for that patient in the long run. N-of-1 trials were introduced to clinicians by Hogben and Sim as early as 1953, but it took 30 years before Gordon Guyatt brought them into the medical mainstream. Early pioneers established active N-of-1 trial units in academic centers, only to abandon them once funding was exhausted. However, several units are still thriving, and over the past three decades, over 2,000 patients have participated in published N-of-1 trials.

And yet, considering the significant potential N-of-1 trials have for individualizing care and supporting shared decision making, a compelling case could be made that they are woefully underused. One reason is that few clinical investigators and even fewer clinicians understand their rationale, methods, and applications. Now, here in one place is the information these individuals have been seeking.

The *Essential Guide to N-of-1 Trials in Health* will be useful to two audiences: clinical researchers seeking a more direct approach to estimating individual treatment effects and clinicians aspiring to apply more rigor to their own therapeutic decision making. Written by many of the world's most knowledgeable authorities on N-of-1 trials, the book provides a step-by-step approach to design and conduct of N-of-1 trials, from study conception and ethical approval to data collection, analysis, and interpretation. While some enthusiastic readers will read the guide cover to cover, each chapter can also stand alone.

When clinicians and patients first hear about N-of-1 trials, their initial incredulity frequently turns to intense interest. How, in an era when personalized medicine is all the rage, could such a powerful approach be so little known? As an accessible yet rigorous introduction to the method, the *Essential Guide to N-of-1 Trials in Health* will help provide tools, answers, and inspiration.

Sacramento, California, USA Richard Kravitz

Acknowledgments

We are very appreciative of the time and skill contributed by Kylie O'Toole and Genevieve Clark in preparing this book for publication and for invaluable administrative support.

We thank our editor at Springer, Melania Ruiz, for her guidance and patience in the process of bringing this book to completion.

Contents

Editors' Biography

Jane Nikles has been working at The University of Queensland in the field of N-of-1 trials for over 15 years. Her Ph.D. on using N-of-1 trial methodology in clinical practice was awarded in 2006. She has been a chief investigator on over $40 m of research funding in the area of N-of-1 trials and has published over 20 peer-reviewed journal articles in the field. She was involved in developing the CONSORT extension for N-of-1 trials (CENT). She is currently conducting an international multisite N-of-1 trial which also compares aggregated N-of-1 trials with parallel arm randomized controlled trials, a world first.

Geoffrey Mitchell is professor of general practice and palliative care at The University of Queensland. His main research interest is in the role of general practitioners in complex conditions, particularly palliative care, and how specialists and GPs can work better together. Current research includes interventions to improve outcomes for caregivers with advanced cancer, health services research in palliative care and primary care, and research into all aspects of N-of-1 trials. He has published over 150 peer-reviewed publications. He has been a chief investigator on over $20 m of research funding. He maintains a clinical general practice in Ipswich, Queensland, Australia, within which he has conducted N-of-1 trials with his patients.

Chapter 1
Introduction

Jane Nikles and Geoffrey Mitchell

N-of-1 trials are multi-cycle within-patient, randomized, double-blind, cross-over comparisons of a drug and placebo (or another drug) using standardized measures of effect. They provide evidence-based information on individual response to treatment and can be used to optimize the chronic disease management of the individual.

Why This Book? Why Now?

With the rising cost of patient care (including drug costs and clinic visits), N-of-1 trials have potential to minimize clinician and patient investment in time and money on suboptimal treatments. Recognition that the USA is in the midst of a healthcare crisis has prompted calls for advances in biomedical research. Potential ways forward are individualized medicine and personalized evidence-based medicine to improve treatment efficiency, by reducing individual patients' exposure to treatments that do not work and those that cause adverse side effects. In addition, moving towards a more individualized and personalized health-care system of the type built from the N-of-1 study principle and infrastructure, would allow exploration and tapping into the potential of genomics and wireless devices. In this context, a text setting out the theoretical and practical issues surrounding N-of-1 trials in the health setting is timely. This is illustrated by a quote from Lillie et al. 2011:

> Despite their obvious appeal and wide use in educational settings, N-of-1 trials have been used sparingly in medical and general clinical settings. We emphasize the great utility of

J. Nikles (✉) • G. Mitchell
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: uqjnikle@uq.edu.au; g.mitchell@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_1

> modern wireless medical monitoring devices in their execution. We ultimately argue that N-of-1 trials demand serious attention among the health research and clinical care communities given the contemporary focus on individualized medicine. (Lillie et al. 2011)

This book presents a comprehensive compendium of issues around the design, conduct, implementation and interpretation of N-of-1 trials in a health system. The contributors are all experts in their own fields as they relate to N-of-1 trials or in N-of-1 trials themselves.

How This Book Came About

Our centre has conducted over 600 N-of-1 trials in areas ranging from osteoarthritis in adults to Attention Deficit Hyperactivity Disorder in children to palliative care. We have experience in conducting the trials face to face and by post and telephone; and both individually and aggregated together.

Our colleagues felt that the expertise we had developed in over 15 years of conducting N-of-1 trials was worth sharing more broadly and more in-depth than is possible in journal articles. The idea for a book was born.

At the time we commenced writing, there were no in depth books on N-of-1 trials in the health setting such as this one. However, Kravitz et al. (2014) have recently published a comprehensive text entitled “Design and Implementation of N-of-1 Trials: A User’s Guide.” Our book intentionally avoids significant overlap with their book.

How to Use This Book

The readers we hope to reach with this book are clinicians, academic researchers, health professionals or practitioners, scientists, and pharmaceutical company staff in the broad area of health; and funders and regulators in various countries who wish to investigate or conduct N-of-1 trials.

The book may also be useful for graduate students in methodologically based courses or doing research higher degrees in areas such as public health, and also for undergraduate students or interested consumers not trained in the health sphere.

We have written this book with two discrete audiences in mind. The first is interested clinicians who will gain benefit from an overview of the N-of-1 technique. We have included chapters that look at the clinical applicability of the technique, how to run an N-of-1 trial in individuals and how to combine results to gain a population estimate. We would suggest reading Chaps. 2, 3, 4, 5, 9, and 15 for this broader overview.

For those readers who desire in-depth examination of N-of-1 trial design, conduct and analysis, we have included chapters that are more technical in nature. This will be of considerable use to people designing high quality trials, and analyzing the data that arises from them, both in terms of determining individual treatment effects and when aggregating the data to generate a population estimate. We would suggest reading Chaps. 6, 7, 8, 10, 11, 12, 13, 14, 16, and 17 for this more in-depth discussion.

A brief description of the chapters follows.

What are N-of-1 trials? In Chap. 2, Jane Nikles defines N-of-1 trials and provides a brief historical perspective. She discusses the background and rationale for N-of-1 trials, and describes their benefits.

Robyn Tate and Michael Perdices' chapter on N-of-1 trials in the behavioral sciences (Chap. 3) describes the application of N-of-1 trials in the behavioural sciences, where they are commonly referred to as single-case experimental designs (SCEDs). Four essential features demarcate single-case methodology from between-group designs: (i) the individual serves as his or her own control, (ii) use of a specific and operationally-defined behaviour that is targeted by the intervention, (iii) frequent and repeated measurement of the target behaviour throughout all phases of the experiment, and (iv) issues surrounding external validity. Features that strengthen internal and external validity of SCEDs are discussed in the context of a standardised scale to evaluate the scientific quality of SCEDs and N-of-1 trials, the Risk of Bias in N-of-1 Trials (RoBiNT) Scale. New work in developing a reporting guide in the CONSORT tradition (the Single-Case Reporting guideline In BEhavioural interventions, SCRIBE) is referenced. Subsequent sections in the chapter highlight differences among the prototypical single-case designs reported in the literature, both experimental (withdrawal/reversal, multiple-baseline, alternating-treatments, and changing-criterion designs) and non-experimental (biphasic A-B design, B-phase training study, preintervention/ post-intervention design, and case description/report), along with illustrative examples reported in the literature. The final section of the chapter describes available methods to analyse data produced by SCEDs, including structured visual analysis, randomization tests and other statistical procedures.

Following on from this, Geoff Mitchell in N-of-1 trials in medical contexts (Chap. 4) argues the case for N-of-1 studies assuming a place in the clinical armamentarium. Clinicians make treatment decisions on a regular basis, and some decisions may result in patients taking treatments for years. This decision-making is a core skill of clinicians, and if possible it should be evidence based. The problem is that the most common tool to aid this decision making, the RCT, has many problems which can lead to a patient being prescribed a treatment that may not work for them. N-of-1 studies may be useful tools to assist in making the best decision possible. This chapter argues the case for N-of-1 studies assuming a place in the clinical armamentarium. It describes the rationale for and uses of N-of-1 trials, the advantages and limitations of N-of-1 trials, and discusses aggregation of N-of-1 trials to generate population estimates of effect.

In the next chapter (Chap. 5) he outlines the rationale, methods, benefits and limitations of combining N-of-1 trials. The original purpose of N-of-1 trials is to determine whether a treatment works in a person. However, these trials can be considered as mini-randomized controlled trials (RCTs), with the person providing multiple datasets to the intervention and control groups. Therefore, several people undergoing the same N-of-1 trial can contribute many data sets and this rapidly scales up to the point where the power of the trial can equate to a normal RCT, but with far fewer participants. This characteristic means that RCT-level evidence can be derived from populations that are almost impossible to gather data from, because of low prevalence conditions, or difficulty in recruiting or retaining subjects. This chapter describes the method in detail, along with methodological challenges and limitations of the method.

Chapter 6 on major design elements of N-of-1 trials by Kimmie Carriere, Yin Li, Geoff Mitchell and Hugh Senior discuss some important considerations when choosing a particular individual N-of-1 trial design. N-of-1 trials are extremely useful in subject-focused investigations, for example, medical experiments. As far as we are aware, no guidelines are available in the literature on how to plan such a trial optimally. In this chapter, they discuss the considerations when choosing a particular N-of-1 trial design, assuming that the outcome of interest is measured on a continuous scale. The discussion is limited to comparisons of two treatments, without implying that the designs constructed can apply to non-continuous or binary outcomes. Optimal N-of-1 trials under various models are constructed depending upon how we accommodate the carryover effects and the error structures for the repeated measurements. Overall, they conclude that alternating between AB and BA pairs in subsequent cycles will result in practically optimal N-of-1 trials for a single patient, under all the models considered, without the need to guess at the correlation structure or conduct a pilot study. Alternating between AB and BA pairs in a single trial is nearly robust to misspecification of the error structure of the repeated measurements.

In Chap. 7 Hugh Senior discusses a major concern in N-of-1 trials, common to any epidemiological approach – the introduction of bias and confounding. These factors may modify the size of the treatment estimate or shift the treatment estimate away from its true value. The methodological approaches of randomization, allocation concealment, and blinding are employed to prevent or minimize confounding and bias in trials. This chapter provides definitions and describes the various methods of randomization, allocation concealment, and blinding that can be adopted in N-of-1 trials. In addition, the chapter details the roles of specific research staff and the information required for the reporting of N-of-1 trial blinding methods in medical journals.

In Chap. 8 on data collection and quality control, Hugh Senior explains how to achieve a reliable data set for analysis that complies with the protocol. A system of clinical data management (the planning and process of data collection, integration and validation) is critical. This chapter provides a synopsis of the key components of clinical data management which need to be considered during the design phase of any trial. Topics addressed include the roles and responsibilities of research staff, the design of case report forms for collecting data; the design and development of a clinical database management system, subject enrolment and data entry, data validation, medical coding, database close-out, data lock and archiving. An additional section discusses the rationale for the requirement of trial registration.

Chapter 9, by Michael Yelland, offers a very practical account of the reporting of N-of-1 trials to patients and clinicians, using trials for chronic pain conditions as models which may be applied to many other forms of N-of-1 trials. It draws from the author's experience in managing N-of-1 trials comparing celecoxib with extended release paracetamol for chronic pain and osteoarthritis and comparing gabapentin with placebo for chronic neuropathic pain.· Reporting the results of N-of-1 trials to patients and health care professionals requires considerable planning to make reports user-friendly and an efficient tool for clinical decision making. Decisions need to be made about key elements of the report, how to order them with the most important summary elements first followed by detailed results, and how to set thresholds for clinically important changes. The inclusion of tables and graphs in reports should improve readability. An example of an individual report is provided.

Adverse events are covered by Hugh Senior in Chap. 10. The safety of subjects who volunteer to participate in clinical trials is paramount. ICH-Good Clinical Practice (ICH-GCP) guidelines assert that 'the rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society'. This chapter describes the internationally accepted standard of the (ICH-GCP) guidelines. It introduces important clinical research terminology, and provides definitions of various types of adverse events, describes the roles and responsibilities of investigators and sponsors, and the processes needed to promote safety through the assessment, recording, evaluating and reporting of adverse events during the design and conduct of clinical trials.

Chapter 11, Research ethics and N-of-1 trials, by Andrew Crowden, Gordon Guyatt, Nikola Stepanov and Sunita Vohra, is an exploration of the ethics of N-of-1 trials and the nature of the relationship between clinical care and clinical research. Some N-of-1 trials are conducted as part of clinical care, others are developed as research. For those that are research, unless they are deemed exempt from formal review, a relevant Human Research Ethics Committee or Institutional Review Board should review specific projects before they are approved. N-of-1 trials should also be authorized by institutions before commencing. The level of risk to the patient/participant should guide and determine whether a particular project is exempt from review, subject to a low/negligible risk review, or should be reviewed by a full committee. Research ethics reviewers must develop a heightened ethical sensitivity toward ensuring that a misguided approach to N-of-1 review does not occur. Clinical researchers, institutions and research review committees, should recognize the continuum of clinical care and clinical research, in order to set and act from explicit standards which are consistent with the clinical practice – clinical research continuum.

Chapter 12 (Kerrie Mengerson, James McGree and Chris Schmid) discusses some techniques for exploratory data analysis and statistical modeling of data from N-of-1 trials, and provides illustrations of how statistical models and corresponding analyses can be developed for the more common designs encountered in N-of-1 trials. Models and corresponding analyses for other designs, perhaps involving different nesting of treatments, order and blocks, can be developed in a similar manner. The focus of this chapter is on continuous response outcomes, that is, numerical response data. The chapter is presented in tutorial style, with concomitant R code and output provided to complement the description of the models. Mixed effects models are also discussed. Such models can be extended to account for a variety of factors whose effects can be considered as random draws from a population of effects. A taxonomy of relevant statistical methods is also presented. This chapter is aimed at readers with some background in statistics who are considering an analysis of data from an N-of-1 trial in the R package.

The economics of N-of-1 trials, Chap. 13, is written by Jennifer Whitty, Joshua Byrnes, and Paul Scuffham, who provide the rationale, challenges and methodological considerations for evaluating the economics of N-of-1 trials. First, they outline the rationale for undertaking an economic evaluation alongside an N-of-1 trial, by describing two key economic questions that are likely to be of interest to researchers, policy makers and clinicians. Then they outline the methods for undertaking an economic evaluation, highlighting some methodological aspects that are of particular relevance for the economics of N-of-1 trials as opposed to more traditional clinical trials. Finally, they acknowledge that the economic evaluation of N-of-1 trials is still in its infancy. We reflect on the research agenda to further develop the potential for N-of-1 trials to inform optimal decision-making around treatment and the appropriate allocation of health care resources.

Next, in Chap. 14, Margaret Sampson, Larissa Shamseer, and Sunita Vohra consider how to describe the individual and aggregated symptom data of N-of-1 trials for professional audiences. Whether an N-of-1 trial is undertaken to inform a particular clinical decision or to test a hypothesis, publishing it in the professional literature may inform other clinical decisions and contribute to the research evidence base. A well-reported N-of-1 trial will provide the transparency needed for readers to critically appraise the work and determine if it is applicable to their situation. A well reported trial can be replicated and, once replicated, results can be aggregated to provide stronger and more compelling evidence. The chapter describes in detail a reporting guideline for N-of-1 trials, CENT (**C**onsort **E**xtension for reporting **N-of-1 T**rials). CENT provides a structured format to ensure that the main journal report is sufficiently detailed that it can be critically appraised and replicated. As well, prospective registration of the trial and data deposit is discussed as means to further increase the transparency and completeness of reporting.

Single Patient Open Trials (SPOTs) are described by Jane Smith, Michael Yelland and Chris Del Mar (Chap. 15). Single patient open trials (SPOTs) are nearly identical to standard trials of treatment. The added essential ingredient is a (commonly arrived at) set of symptoms to monitor (the *outcome measure*). This means they lie somewhere in between formal N-of-1 trials and totally informal trials of treatment in terms of rigour. SPOTs are accordingly less demanding to arrange (for both the patient and clinician) than N-of-1 trials, but they require considerably more effort and commitment than casual trials of treatment. This chapter defines and describes the rationale for SPOTs, discusses when and why they could be used, as well as their limitations, and describes outcome measures and analysis. As well as describing the use of SPOTs in clinical contexts, it covers the extra considerations required when using SPOTs in research. Several examples of the practical application of SPOTs are given, some with the resulting data. It is anticipated that the examples may be adapted to enable other clinicians and their patients to perform their own SPOTs to validate other medical interventions in the context of the individual.

Next, Kerrie Mengersen, James McGree and Christopher Schmid discuss issues and approaches related to systematic review and meta-analysis of N-of-1 trials. Chapter 16 describes some basic guidelines and methods, and some important steps in a systematic review of these types of trials are discussed in detail. This is followed by a detailed description of meta-analytic methods, spanning both frequentist and Bayesian techniques. A previously undertaken meta-analysis of a comparison of treatments for fibromyalgia syndrome is discussed with some sample size considerations. This is further elaborated on through a discussion on the statistical power of studies through a comparison of treatments for chronic pain. The chapter concludes with some final thoughts about the aggregation of evidence from individual N-of-1 trials.

Finally, in Chap. 17, Jane Nikles looks at the current status of N-of-1 trials and where N-of-1 trials are headed. N-of-1 trials and review articles have recently been published in the areas of chronic pain, pediatrics, palliative care, complementary and alternative medicine, rare diseases, patient-centered care, the behavioral sciences and genomics. These are briefly reviewed and the current place of N-of-1 trials discussed. The chapter concludes with a vision for the future of N-of-1 trials.

We trust you find the book useful. Feedback that might inform later editions is welcomed.

References

Kravitz RL, Duan N (eds) and the DEcIDE Methods Center N-of-1 Guidance Panel (Duan N, Eslick I, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S). Design and implementation of N-of-1 trials: a user's guide. AHRQ Publication No. 13(14)-EHC122-EF. Agency for Healthcare Research and Quality, Rockville, February 2014. www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm

Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ (2011) The N-of-1 clinical trial: the ultimate strategy for individualizing medicine? Pers Med 8(2):161–173

Chapter 2
What are N-of-1 Trials?

Jane Nikles

Abstract In this chapter, we define N-of-1 trials and provide a brief historical perspective. We briefly cover the background and rationale for N-of-1 trials, and discuss their benefits.

Keywords N-of-1 trials • Trial of therapy • Clinical trial • Crossover trial • Patient-centered outcome research • Chronic-disease • Medication • Medication expenditure • Quality use of medicines • Adverse events

Introduction

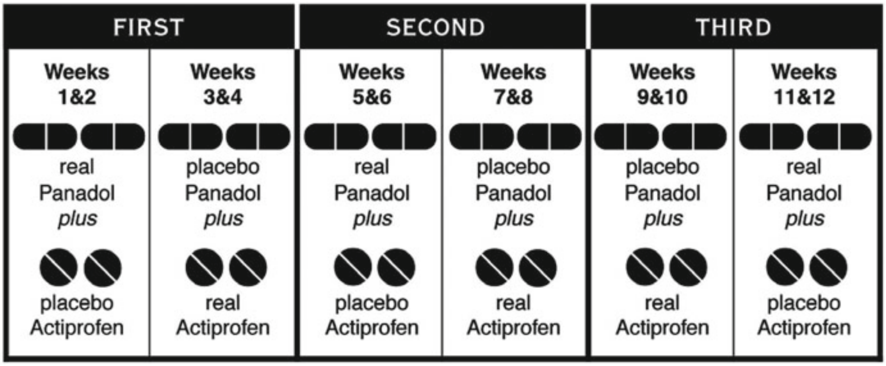
N-of-1 trials are within-patient, randomized, double-blind, cross-over comparisons of a treatment with either a placebo or another treatment. Because the trials involve only one patient, with that patient acting as their own control, the comparison is made between multiple (usually three periods for each alternative) cross-over time intervals.

N-of-1 trials can identify those individual patients who respond (and those who do not) to particular medications for chronic, stable conditions. For example, the trials allow the identification of patients who obtain a satisfactory response from cheaper or safer drugs (for example, paracetamol) and those who might gain additional benefit from more expensive, less well-tolerated drugs, for instance non-steroidal anti-inflammatory drugs (NSAIDs).

A classic example of a patient who responded was a 57-year-old woman who had a long history of knee osteoarthritis (OA) symptoms not adequately controlled on regular paracetamol. She underwent an N-of-1 trial of ibuprofen 400 mg tds versus paracetamol 1 g tds (Fig. 2.1). Pain and stiffness visual analog score (VAS) plots showed a clear and significant advantage of the NSAID over paracetamol, with the patient also preferring the NSAID at two of the three changeovers. This woman

J. Nikles (✉)
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: uqjnikle@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_2



FIRST		SECOND		THIRD	
Weeks 1&2	Weeks 3&4	Weeks 5&6	Weeks 7&8	Weeks 9&10	Weeks 11&12
real Panadol *plus*	placebo Panadol *plus*	real Panadol *plus*	placebo Panadol *plus*	placebo Panadol *plus*	real Panadol *plus*
placebo Actiprofen	real Actiprofen	placebo Actiprofen	real Actiprofen	real Actiprofen	placebo Actiprofen

Fig. 2.1 Example arrangement of treatment order

clearly responded to NSAIDs, and was subsequently prescribed ketoprofen, 200 mg daily (Nikles et al. 2005).

N-of-1 trials can be used in many different ways; for example:

1. To determine whether a certain drug should be started
2. A drug to test whether medication can be discontinued
3. To compare different dosages of the same drug
4. To compare different brands of the same drug, or
5. To ascertain whether certain symptoms are side effects of a particular drug.

N-of-1 trials are inherently best suited for the following (Guyatt et al. 2000):

1. Areas that lack evidence for individual patients
2. Therapies that have varying effects across patients (e.g., anticonvulsants for neuropathic pain)
3. Stable or chronic conditions to maximize knowledge gained from N-of-1 trials for future clinical decisions (e.g., gastroesophageal reflux disease)
4. Treatments that have a rapid onset of effect (i.e., short half--life)
5. Minimal "washout" periods (i.e., time needed for one treatment to dissipate and the next to initiate and stabilize)
6. Stability of treatment to ensure that information gained from N-of-1 trials can adequately direct future clinical decisions
7. When there are validated measures for treatment effects.

Additionally, the question must be important – either because of cost, serious potential adverse effects, or clinical or societal importance.

They do not work when:

1. The outcome is prevention of an event over a long period of time
2. Treatment alters the underlying condition (e.g., testing antibiotics), or
3. Treatment effects are cumulative or extended over long periods of time (e.g., antidepressants for depression).

These trials may also be useful in situations where there is a need to prioritize medications or where there is significant difference in cost or availability between drugs approved for the same indication.

In addition, for selected treatments, multiple N-of-1 studies of the same treatment in a similar patient population can be aggregated with high levels of power (e.g. via Bayesian or other statistical methods), to provide a population estimate of effect, but requiring a fraction of the sample size of the equivalent parallel arm RCT. This has obvious benefits for accumulating evidence in populations where participants are hard to recruit (e.g., small number of patients such as pediatric traumatic brain injury (Nikles et al. 2014), rare conditions (Facey et al. 2014) or to retain (e.g., palliative care (Mitchell et al. 2015). Further, the effect on each participant in the trial is known, which presents opportunities for research into individual responsiveness to therapy. This is discussed in more detail in Chap. 5.

Historical Perspective

Cushny and Peebles conducted the first N-of-1 trial in 1905 by examining the saliva--inducing effects of optical isomers of hyoscines (Cushny and Peebles 1905). N-of-1 trials have long been used in psychology, and over the last 40 years, have been used in many different situations in clinical medicine, initially by Kellner in 1968 (Kellner and Sheffield 1968) as an experimental approach to targeting drug therapy. A resurgence in 1986 with Guyatt, Sackett and colleagues conducting several N-of-1 trials and the setting up of N-of-1 services by Guyatt and Larson in the 1980s and 1990s (e.g. Guyatt et al. 1986, 1988; Larson 1990; Larson et al. 1993), marked the beginning of the use of N-of-1 trials in modern clinical settings.

Background and Rationale

High Burden of Chronic Disease in Western Society

Today, chronic diseases are among the most prevalent, costly, and preventable of all health problems. Seven out of every ten Americans who die each year, or more than 1.7 million people in America yearly, die of a chronic disease (Centers for Chronic Disease Prevention and Health Promotion 2009). The prolonged course of illness and disability from such chronic diseases as diabetes and arthritis results in extended pain and suffering and decreased quality of life for millions of Americans, causing major limitations in activity for more than 33 million Americans (Centers for Chronic Disease Prevention and Health Promotion 2009).

High Cost of Medication for Chronic Disease

In the USA, as in other Western countries, chronic diseases cost large amounts both to the government and to the individual. Heart disease and stroke cost US$313.8 billion in 2009; cancer cost US$89 billion in health care expenditures in 2007; the direct costs of diabetes totaled $US116 billion in 2007; the cost of direct medical care for arthritis was more than $US80.8 billion per year, in 2003 (Centers for Chronic Disease Prevention and Health Promotion 2009).

In Western countries, a large part of the cost of chronic disease is the cost of medications and their adverse effects. As in the USA, large and increasing amounts of money are spent on prescription medications in Australia. In the year ending 30 June 2013, a total of at least AUD$8.9 billion, including AUD$7.4 billion by the PBS and AUD$1.49 billion by patients, was spent on 197 million PBS prescriptions (compared to 194.9 million in the previous year) (PBS expenditure and prescriptions, July 2012).

Inappropriate Medication Use

Over the last 15 years, there has been increasing concern about the high psychosocial, economic and health costs of inappropriate medication use. Many people do not respond to medicines they are prescribed (Guyatt et al. 2000): the evidence that clinicians rely on to make treatment decisions may not apply to individuals. Ensuring that patients only take medicines that work for them, is an important strategy in reducing the burden of medication misadventure. In Australia each year, medication misadventure, including adverse drug reactions and drug-drug interactions, is implicated in 2–3 % of all hospitalizations (i.e. 190,000 per year) (Australian Council for Safety and Quality in Health Care (2006)), 8,000 hospital related deaths (Roughead and Semple 2009), an estimated annual cost of $350 million in direct hospital costs and total costs to the health system of $660 million (Roughead and Semple 2009). Even for those who are not ill enough to require admission, adverse events can impair their quality of life (Sorensen et al. 2005). Polypharmacy, especially in older people and women, contributes to this (Curry et al. 2005). There is also evidence to suggest that considerable amounts of medicines are wasted, that is, unused by patients and returned to pharmacists (Commonwealth Department of Human Services and Health 2009).

The Need to Find Ways of Targeting Medications

It is clear that the appropriate and safe use of medicines is an urgent national priority. In the current evidence-based and consumer-driven policy environments, information which helps medical practitioners and patients make informed decisions about the appropriate use of medicines is much needed, for example, ways of reducing unnecessary medication use by targeting medications to responders.

Although the art of prescribing has advanced significantly over recent decades, there is still a large element of uncertainty involved. For example, it is known that only 30–50 % of individuals with osteoarthritis respond to NSAIDs (Walker et al. 1997); so how does a clinician predict whether a particular osteoarthritic patient will respond or not?

Identifying Patients Who Respond

Genetic variation or polymorphism may be an important factor underlying the variation in individual responses to certain drugs. For example, genetic defects in the dopamine transporter gene might contribute to some forms of ADHD; thus explaining why certain individuals respond to psychostimulants, which interact with the dopamine, serotonin, and norepinephrine transporters of monoamines across the plasma membrane (Jayanthi and Ramamoorthy 2005)**.**

Until pharmacogenetics – the study of human genome function and its effects on drug response – becomes further developed and widely available, N-of-1 trials remain the best method of identifying patients who respond to certain drugs.

Results of Clinical Trials Not Necessarily Applicable to Individuals

Randomized controlled trials remain the ‘gold standard’ for assessing drug effectiveness. However, the difficulty of extrapolating the results of randomized controlled trials to the care of an individual patient has resulted in prescribing decisions in clinical practice often being based on tradition, rather than evidence (Larson 1990; Larson et al. 1993; Mapel et al. 2004). When considering any source of evidence about treatment other than N-of-1 trials, clinicians are generalizing from results on other people to their patients, inevitably weakening the inferences about treatment impact and introducing complex issues of how randomized controlled trial (RCT) results apply to individuals (Kellner and Sheffield 1968). In fact, in their hierarchy of study design to evaluate the strength of evidence for making treatment decisions, Guyatt et al. place N-of-1 trials at the top (highest level of evidence) (Guyatt et al. 2000). This is discussed in more detail in Chap. 4.

How do Doctors Decide on Medications at Present? The Trial of Therapy

Doctors often use a trial of therapy to assist in their clinical decision-making. That is, the patient presents with a particular cluster of symptoms, they are prescribed a particular medication (for example, an asthma drug or an NSAID) tentatively, as a trial, and the subsequent condition of the patient is used to monitor the efficacy of the treatment, usually informally. Then, based on the patient's response, the medication is either continued, discontinued or the dose is changed. This is discussed in more detail in Chap. 4.

Biases of Informal Trials of Therapy

The informal trial of therapy has serious potential biases. These are the placebo effect, the patient's desire to please the doctor, and the expectations of patient and doctor. Although commonly used, it is not adequate for determining the appropriateness of prescribing certain medications, particularly those that have significant side effects or are expensive. This is discussed in more detail in Chap. 4.

N-of-1 Trials as a Possible Solution

One method with the potential to solve these problems is the N-of-1 trial. An N-of-1 trial is more rigorous and objective version of a trial of therapy and therefore could be a potentially feasible initiative to incorporate into clinicians' routine day to day practice. Rather than using a *group* of patients as a control for a group of patients, as is done in parallel group clinical trials, N-of-1 trials use a single patient as their own control. Because the data have come from that patient, the result is definitely applicable to that patient. Because treatment periods are randomized and double-blind, N-of-1 trials remove the biases mentioned above.

N-of-1 Trials Are Widely Applicable

The N-of-1 trial is a more sophisticated refinement of the trial of therapy. Clinicians could, potentially, adopt this clinical tool on a routine basis to assist in medication-related decision-making. A recent review found 108 N-of-1 trials for medications, medical devices, surgical treatments, acute conditions, and behavioral interventions since 1986 (Gabler et al. 2011). The most common conditions examined in the N-of-1 trials were neuropsychiatric (36 %, of which 9 % were attention deficit hyperactivity disorder), musculoskeletal (21 %, of which 9 % were osteoarthritis), and pulmonary (13 %).

Potential Benefits of N-of-1 Trials

Patients

Participation in their clinical management is a very powerful tool in enabling the individual to make informed therapeutic choices. This is one of the principles of patient care. Patients may benefit from involvement in N-of-1 trials because they can have a trial conducted for their own situation, in which they can personally determine the benefits of two therapeutic choices (Nikles et al. 2005).

Because of involvements in the data collection and decision making processes, N-of-1 trials may increase patients' confidence in, and commitment to, their long-term pharmacological management (Nikles et al. 2005). Patients who do not respond could avoid any possible side effects from taking a medication that is not effective. Targeting therapy to only those who benefit could have a significant positive impact on health outcomes for patients with chronic conditions.

Doctors

The pharmacological management of various chronic stable conditions such as OA, in which individual response to treatment is variable, forms a large part of many medical practitioners' workload. Improving the appropriateness and precision of prescribing for these conditions, by identifying those who respond to particular drugs and those who do not, could make a significant contribution to improving the quality of patient care. N-of-1 trials may also aid the further development of trust and mutual respect within the doctor-patient therapeutic partnership.

Using N-of-1 trial methods may encourage medical practitioners to apply rigorous methods to evaluate both standard and new therapies. This could be important in developing critical appraisal skills, and empowering to those medical practitioners who have felt that research was not of direct use for resolving their problems and questions (Askew 2005). Providing non-academic medical practitioners with an appreciation of the benefits of research is much needed (Moulding et al. 1997). N-of-1 trials may help doctors to realize the effectiveness of certain drugs in their individual patients in a way that they are not able to appreciate otherwise. This could be a very powerful educating influence. N-of-1 trials may also influence prescribing indirectly by encouraging more thoughtful prescribing of medications in general.

Health System

Anticipated benefits include: the generation of innovative treatment approaches for individual patients, more appropriate and cost effective prescribing, the minimization of the risk of adverse side effects, reduced inappropriate and wasteful prescribing, and therefore decreased unnecessary expenditure. Cost savings could benefit the PBS, the Repatriation Pharmaceutical Benefits Scheme (RPBS) – this is the PBS equivalent for war veterans and their families – and private health funds, which provide set levels of reimbursement for certain medications and patients.

N-of-1 trials in the behavioral sciences will be further explored in Chap. 3 and N-of-1 trials in medical contexts in Chap. 4.

Conclusion

N-of-1 trials are individual randomized multiple crossover controlled trials which use the patient as their own control. Applicable in many contexts, they have significant benefits for patients, clinicians and the health system.

References

Askew DA (2005). A study of research adequacy in Australian general practice. PhD thesis, The University of Queensland, Brisbane

Australian Council for Safety and Quality in Health Care (2006) National inpatient medication chart: general instructions/information for doctors. The Australian Department of Health and Ageing, Canberra

Centers for Chronic Disease Prevention and Health Promotion (2009) The power of prevention chronic disease….the public health challenge of the 21st century. http://www.cdc.gov/chronicdisease/pdf/2009-power-of-prevention.pdf. Accessed 31 Oct 2014

Commonwealth Department of Human Services and Health (2009) A policy on the quality use of medicines. Commonwealth Department of Human Services and Health, Canberra

Curry LC, Walker C, Hogstel MO, Burns P (2005) Teaching older adults to self manage medications: preventing adverse drug reactions. J Gerontol Nurs 31(4):32–42

Cushny AR, Peebles AR (1905) The action of optical isomers: II. Hyoscines. J Physiol 32(5–6):501–510

Facey K, Granados A, Guyatt G, Kent A, Shah N, van der Wilt GJ, Wong-Rieger D (2014) Generating health technology assessment evidence for rare diseases. Int J Technol Assess Health Care 19:1–7 [Epub ahead of print]

Gabler NB, Duan N, Vohra S, Kravitz RL (2011) N-of-1 trials in the medical literature: a systematic review. Med Care 49(8):761–768. doi:10.1097/MLR.0b013e318215d90d

Guyatt G, Sackett D, Taylor DW et al (1986) Determining optimal therapy: randomized trials in individual patients. N Engl J Med 314:889–892

Guyatt G, Sackett D, Adachi J et al (1988) A clinician's guide for conducting randomized trials in individual patients. CMAJ 139:497–503

Guyatt GH et al (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

Jayanthi LD, Ramamoorthy S (2005) Regulation of monoamine transporters: influence of psychostimulants and therapeutic antidepressants. AAPS J 7(3):E728–E738

Kellner R, Sheffield BF (1968) The use of self-rating scales in a single-patient multiple cross-over trial. Br J Psych 114(507):193–196

Larson EB (1990) N-of-1 clinical trials. A technique for improving medical therapeutics. West J Med 152:52–56

Larson EB, Ellsworth AJ, Oas J (1993) Randomised clinical trials in single patients during a 2–year period. JAMA 270(22):2708–2712

Mapel DW, Shainline M, Paez K, Gunter M (2004) Hospital, pharmacy and outpatient costs for osteoarthritis and chronic back pain. J Rheum 31(3):573–583

Mitchell G, Hardy J, Nikles J, Carmont S, Senior H, Schluter P, Good P, Currow D (2015) The effect of methylphenidate on fatigue in advanced cancer: an aggregated n-of-1 trial. J Pain Symptom Manage. doi:10.1016/j.jpainsymman.2015.03.009

Moulding N, Fahy N, Heng Foong L, Yeoh J, Silagy C, Weller D (1997) A systematic review of the current status of evidence-based medicine and its potential application to Australian general practice. A report to the Commonwealth Department of Health and Family Services. Department of General Practice, Flinders University

Nikles CJ, Clavarino AM, Del Mar CB (2005) Using N-of-1 trials as a clinical tool to improve prescribing. Br J Gen Pract 55(512):175–180

Nikles J, McKinlay L, Mitchell G, Carmont S-A, Senior H, Waugh M-C, Epps A, Schluter P, Lloyd O (2014) Aggregated N-of-1 trials of CNS stimulants versus placebo for paediatric traumatic brain injury – a pilot study. Trials 15:54

PBS expenditure and prescriptions, July 2012 to June 2013. Pharmaceutical benefits scheme website. http://www.pbs.gov.au/statistics/2012-2013-files/expenditure-and-prescriptions-12-months-to-30-06-2013.pdf. Accessed 31 Oct 2014

Roughead EE, Semple SJ (2009) Medication safety in acute care in Australia: where are we now? Part 1: a review of the extent and causes of medication problems 2002–2008. Australia and New Zealand Health Policy, 6. http://www.anzhealthpolicy.com/content/6/1/18. Accessed 18 Aug 2013

Sorensen L, Stokes JA, Purdie DM, Woodward M, Roberts MS (2005) Medication management at home: medication-related risk factors associated with poor health outcomes. Age Ageing 34(6):626–632

Walker JS, Sheather-Reid RB, Carmody JJ, Vial JH, Day RO (1997) Nonsteroidal anti-inflammatory drugs in rheumatoid arthritis and OA: support for the concept of “responders” and “non-responders”. Art Rheum s40(11):1944–1954

Chapter 3
N-of-1 Trials in the Behavioral Sciences

Robyn L. Tate and Michael Perdices

Abstract This chapter describes the application of N-of-1 trials in the behavioural sciences, where they are commonly referred to as single-case experimental designs (SCEDs). Four essential features demarcate single-case methodology from between-group designs: (i) the individual serves as his or her own control, (ii) use of a specific and operationally-defined behaviour that is targeted by the intervention, (iii) frequent and repeated measurement of the target behaviour throughout all phases of the experiment, and (iv) issues surrounding external validity. Features that strengthen internal and external validity of SCEDs are discussed in the context of a standardised scale to evaluate the scientific quality of SCEDs and N-of-1 trials, the Risk of Bias in N-of-1 Trials (RoBiNT) Scale. New work in developing a reporting guide in the CONSORT tradition (the Single-Case Reporting guideline In BEhavioural interventions, SCRIBE) is referenced. Subsequent sections in the chapter highlight differences among the prototypical single-case designs reported in the literature, both experimental (withdrawal/reversal, multiple-baseline, alternating-treatments, and changing-criterion designs) and non-experimental (biphasic A-B design, B-phase training study, pre-intervention/post-intervention design, and case description/report), along with illustrative examples reported in the literature. The final section of the chapter describes available methods to analyse data produced by SCEDs, including structured visual analysis, randomization tests and other statistical procedures.

Keywords Single-case experimental designs • N-of-1 trials • Behavioral sciences • Design • Analysis • Test interpretation • Single-case methodology • Non-experimental • Case description • B-phase training • Pre-post intervention • Biphasic design • Withdrawal/reversal design • Multiple baseline • Alternate treatment • Changing-criterion design

R.L. Tate (✉)
John Walsh Centre for Rehabilitation Research, Kolling Institute of Medical Research, Sydney Medical School-Northern, University of Sydney, Sydney, Australia
e-mail: rtate@med.usyd.edu.au

M. Perdices
Department of Neurology, Royal North Shore Hospital, Sydney, Australia

Department of Psychological Medicine, Sydney Medical School-Northern, University of Sydney, Sydney, Australia
e-mail: Michael.Perdices@health.nsw.gov.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_3

Evolution of the N-of-1 Trial in the Behavioral Sciences

In the 25 years between 1985 and 2010, just over 100 articles were published in medical journals using N-of-1 methodology (Gabler et al. 2011). By contrast, during the same period more than 600 articles using single-case experimental designs were published in the neuropsychological rehabilitation field alone (www.psycbite.com, accessed 21 February, 2014), and in the even more circumscribed field of special education more than 450 such articles were published over approximately the same period (1983–2007; Hammond and Gast 2010).

The behavioral sciences (including clinical, education, and neuro- and rehabilitation psychology) have used single-case experimental methodology to test the efficacy of interventions for many decades. Indeed, Gordon Guyatt, who was interviewed about the evolution of the medical N-of-1 trial that occurred in the early 1980s, commented that at that time:

> The department [Clinical Epidemiology and Biostatistics at McMaster University, Canada] was multidisciplinary and very tightly integrated. So there were … statisticians and psychologists and people with behavioral backgrounds, physicians and epidemiologists getting together on a regular basis. And for a while, one of the psychologists would say, 'Oh, that would be very interesting for an N-of-1 trial.' And we said, 'Thank you very much' and would go on. Then at one point it clicked, and we started to get out the psychology literature and found three textbooks full of N-of-1 designs from a psychology perspective. … It was totally old news (Kravitz et al. 2008, p. 535).

This interview highlights two issues: first, the methodology of the N-of-1 trial was already well established in psychology by the time that Guyatt and colleagues commenced work with N-of-1 trials, and second, there was not just a single N-of-1 design but many different types. In what is taken to be the initial, landmark publication describing a medical N-of-1 trial (a 66-year old man with uncontrolled asthma), Guyatt et al. (1986, pp. 889–890) commented that "experimental methods have been applied to individual subjects in experimental psychology for over two decades, to investigate behavioral and pharmacological interventions. The method has been called an 'intensive research design', a 'single case experiment' or (the term we prefer) an 'N of 1' study ('N' being a standard abbreviation for sample size)." In the behavioral sciences, although the term 'N-of-1' was used in early work (e.g., Davidson and Costello 1978; Kratochwill and Brody 1978), the descriptor 'single-case experimental design' (SCED) is more commonly used, referring to a family of different types of controlled research designs using a single participant. Because the 'N-of-1 trial' in medicine now refers to a specific type of SCED, henceforth, in the present chapter we use the broader term SCED to avoid confusion. We include the medical N-of-1 trial as a subset of SCEDs, the varieties of which are described later in this chapter.

Several essential features define SCEDs and distinguish them from traditional between-group research methodology. These differences are important, because they also form the building blocks for designing and implementing scientifically rigorous SCEDs:

First, in SCEDs the individual serves as his or her own control. Whereas a group design compares two or more groups who receive different interventions (which may include a non-intervention condition), comparable control is achieved in SCEDs by having the *same* individual receive the intervention conditions sequentially in a number of "phases" (this term being comparable to "periods" as used in the medical N-of-1 literature). There is a minimum of two types of phases, the baseline phase (generally designated by the letter, A) and the intervention phase (generally designated by the letter, B). In this way, a SCED involves a controlled experiment. There are certain design rules that govern the way in which the intervention (i.e., the independent variable) is manipulated (e.g., only changing one variable at a time) which are described in detail in single-case methodology texts (e.g., Barlow et al. 2009; Kazdin 2011).

Second, the outcome (i.e., the dependent variable) in SCEDs is an operationally defined and precisely specified behavior or symptom that is targeted by the intervention (and hence referred to as the "target behavior"). In group designs, outcomes often reflect general constructs (e.g., social skills) and are usually measured with standardized instruments that, ideally, have good psychometric properties. Such instruments (at least in the behavioral sciences) often encompass multiple and even disparate aspects of the outcome of interest (e.g., an outcome measure for "social skills" may include a heterogeneous set of items addressing eye contact, facial expression, initiating conversation, response acknowledgement) and the score obtained may not capture the specificity of the particular problem behavior being treated. By contrast, the dependent variable used in a SCED aims to provide a specific, precisely defined, behavioral observation (e.g., frequency of initiating conversation topics), and indeed often does not employ a standardized instrument for measuring the outcome. As a consequence, the researcher needs to demonstrate that the data collected on the target behavior have acceptable inter-rater reliability.

Third, SCEDs involve frequent and repeated measurement of the target behavior in every phase, whereas the outcome variable/s in group designs may be measured on as little as two occasions (pre-intervention and post-intervention). The reason for multiple measures of the target behavior is to address the variability in behavior that occurs in a single individual. In group designs, such variability is overcome by aggregating data across participants. Because the target behavior in SCEDs needs to be measured repeatedly, it also needs to lend itself well to this purpose and be amenable to frequently repeated administrations (which are often not feasible with standardized instruments that may be time-consuming to administer). The exact number of measurements taken per phase, however, will also depend on different parameters, such as stability of the data (discussed later in this chapter).

Fourth, another difference between group designs and SCEDs pertains to external validity, specifically generalization of the results to other individuals and other settings. Traditionally, external validity has been regarded as a special strength of group designs (most notably in the randomized controlled trial), and conversely a particular weakness of SCEDs – clearly, with a sample size of n = 1, the grounds for generalization to other individuals are very tenuous. Nonetheless, these extreme views regarding external validity are an oversimplification. It has been observed previously that generalization of results from a group design only applies to individuals who share similar characteristics to those participating in the study, and more specifically, the subset of participants in a study who actually improve (generally, not all participants respond positively to the intervention). Moreover, selection criteria for participants in clinical trials are often stringent in order to increase homogeneity of the sample and generally people are excluded who have premorbid and/or current comorbidities (e.g., in the neuropsychological rehabilitation field: alcohol and substance use problems, other neurological or psychiatric conditions, as well as severe motor-sensory and cognitive impairments). Restricted selection criteria place severe limitations on the capacity to generalize the results to those individuals who do not possess such characteristics.

On the other hand, the SCED has developed methods that strengthen external validity, most commonly through replication (see Horner et al. 2005), as well as using generalization measures as additional outcome measures (see Schlosser and Braun 1994). Direct replication of the experimental effect *within the experiment* (i.e., intra-subject replication) is a key feature of single-case methodology that strengthens internal validity, whereas inter-subject replication and systematic replication are methods to enhance external validity (Barlow et al. 2009; Gast 2010; Horner et al. 2005; Sidman 1960).

The foregoing provides the parameters of single-case methodology. It is important to distinguish this single-case methodology from anecdotal, uncontrolled case descriptions that are also reported in the literature. Specifically, single-case methodology is distinguished by the following cardinal features (Allen et al. 1992; Backman et al. 1997; Perdices and Tate 2009; Rizvi and Nock 2008):

- It consists of a number of discrete phases, generally, but not invariably, baseline (A) and intervention (B) phases in which the individual serves as his or her own control
- There is a clear, operational definition of the dependent variable (behavior targeted for treatment)
- The dependent variable is measured repeatedly and frequently throughout all phases using precisely defined methods
- Measurement and recording procedures are continued until requirements of the specific design have been satisfied
- Experimental control is exercised by systematically manipulating the independent variable (intervention), manipulating one independent variable at a time and carefully controlling extraneous variables.

The above defining features of SCEDs have been incorporated into methodological quality rating scales to evaluate both internal and external validity. It is well established that even randomized controlled trials vary widely with respect to their scientific rigor (in the neurorehabilitation literature see Moseley et al. 2000; Perdices et al. 2006), and the single-case literature also exhibits variability with respect to scientific calibre (Barker et al. 2013; Maggin et al. 2011; Shadish and Sullivan 2011; Smith 2012; Tate et al. 2014). Over the past 10 years concerted efforts have been made in the behavioral sciences to improve the conduct and reporting of SCED

Table 3.1 Item content of the risk of bias in N-of-1 trials (RoBiNT) scale

Internal validity subscale	External validity and interpretation subscale
1. Design: Does the design of the study meet requirements to demonstrate experimental control?	8. Baseline characteristics: Were the participant's relevant demographic and clinical characteristics, as well as characteristics maintaining the condition adequately described?
2. Randomization: Was the phase sequence and/or phase commencement randomized?	9. Therapeutic setting: Were both the specific environment and general location of the investigation adequately described?
3. Sampling: Were there a sufficient number of data points (as defined) in each of baseline and intervention phases?	10. Dependent variable (target behavior): Was the target behavior defined, operationalized, and the method of its measurement adequately described?
4. Blind participants/therapists: Were the participants and therapists blinded to the treatment condition (phase of study)?	11. Independent variable (intervention): Was intervention described in sufficient detail, including the number, duration and periodicity of sessions?
5. Blind assessors: Were assessors blinded to treatment condition (phase of study)?	12. Raw data record: Were the data from the target behavior provided for each session?
6. Inter-rater reliability (IRR): Was IRR adequately conducted for the required proportion of data, and did it reach a sufficiently high level (as defined)?	13. Data analysis: Was a method of data analysis applied and rationale provided for its use?
7. Treatment adherence: Was the intervention delivered in the way it was planned?	14. Replication: Was systematic and/or inter-subject replication incorporated into the design?
	15. Generalization: Were generalization measures taken prior to, during, and at the conclusion of treatment?

Note: the RoBiNT manual, available from the corresponding author, provides detailed description and operational definitions of the items

research. Standards of design and evidence reported by Kratochwill et al. (2013) draw heavily on seminal work in the area published by Horner and colleagues (2005). All of these developments were used by Tate et al. (2013) in revising their methodological rating scale to evaluate the scientific rigor of SCEDs.

The 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale (Tate et al. 2013) examines both internal validity (7 items) and external validity and interpretation (8 items), as shown in Table 3.1. The original impetus to develop the RoBiNT Scale (and its precursor the 11-item SCED Scale; Tate et al. 2008) was to critically appraise the scientific quality of SCEDs in the published literature for the purpose of rating methodological quality of the studies archived on our PsycBITE database.[1]

[1] PsycBITE is a multi-disciplinary database that archives all of the published empirical articles on nonpharmacological treatments for the psychological consequences of acquired brain impairment and contains more than 4,500 records. Controlled studies (both group and single-case) are rated for scientific rigor and ranked on the database in terms of their scientific quality.

The specific requirements of our research team were to develop a practical, feasible, reliable and sensitive scale that evaluated risk of bias in single-case research designs. The RoBiNT Scale is designed to be a generic scale, applicable to any type of SCED (including the medical N-of-1 trial). It demonstrates very good psychometric properties, with excellent inter-rater reliability (all intra-class correlation coefficients (ICC) for each of the total score and the two subscales for pairs of both novice and trained raters are in excess of ICC = 0.86), as well as discriminative validity (Tate et al. 2013).

An added feature of the RoBiNT Scale is that the items can be used as a checklist to plan the conduct and reporting of single-case experiments and N-of-1 trials. That said, recent endeavors have focused specifically on developing reporting guidelines in the CONSORT tradition for SCEDs in the behavioral sciences literature (Tate et al. 2012). The guideline, entitled the SCRIBE (**S**ingle-**C**ase **R**eporting guideline **I**n **BE**havioral interventions) is to be published soon (Tate et al. accepted). The SCRIBE owes its origins to the complementary CONSORT Extension for N-of-1 Trials (CENT) Statement developed for the medical community (Shamseer et al. 2015; Vohra et al. 2015; see Chap. 14 in this volume). The reason that two sets of guidelines are being developed is to cater to the different readership (medical vs behavior sciences). In addition, the CENT relates exclusively to the medical N-of-1 trial, whereas the SCRIBE addresses the broader family of SCEDs.

Common Types of Designs Using a Single Participant

Perusal of reports in the literature describing a single participant reveals a plethora of different types of designs, some of which do not conform to the defining features of single-case methodology listed earlier in this chapter. Based on our survey of the neuropsychological literature reporting on a single participant (Perdices and Tate 2009), along with our work in classifying more than 1,000 published single-case reports archived on PsycBITE, we have mapped the common types of designs using individualized data from one or several individual participants. Figure 3.1 which is taken from our in-house training manual for rating scientific rigor of single-case designs, was informed by discussion at the CENT consensus conference in May 2009, and is slightly adapted from the figure published in Tate et al. (2013) so that it specifies the location of the N-of-1 trial. Designs that use a single participant can divided into two main classes:

(a) those using single-case methodology, some of which involve experimental control (depicted above the *horizontal line* in Fig. 3.1), and others which do not have experimental control (viz., the bi-phasic A-B design).
(b) non-experimental designs, which are shown below the *horizontal line*.

The main varieties of SCEDs are described in the following section. In terms of non-experimental designs, these comprise case descriptions, single phase (B-phase

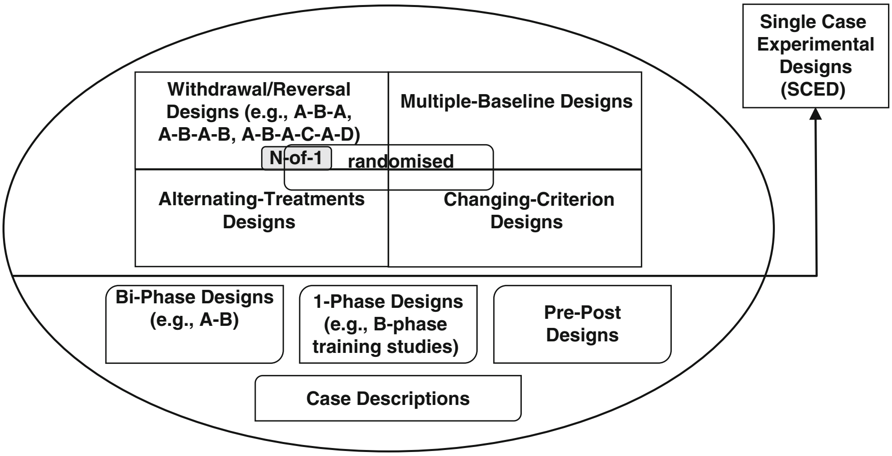

Single Case Experimental Designs (SCED)
Withdrawal/Reversal Designs (e.g., A-B-A, A-B-A-B, A-B-A-C-A-D)
Multiple-Baseline Designs
N-of-1
randomised
Alternating-Treatments Designs
Changing-Criterion Designs
Bi-Phase Designs (e.g., A-B)
1-Phase Designs (e.g., B-phase training studies)
Pre-Post Designs
Case Descriptions

Fig. 3.1 Common types of designs using a single participant (Reproduced from: unpublished manual for critical appraisal of single-case reports, University of Sydney. Adapted version published in Tate et al. (2013))

training) studies, pre-intervention/post-intervention designs, along with the bi-phasic (A-B) design.

Case Descriptions

Case descriptions are "a description of clinical practice that does not involve research methodology" (Backman and Harris 1999, p. 171). The quintessential example of the case description noted by Rizvi and Nock (2008) is the classic work of Freud and Breuer (1895), the most famous being the case of Anna O.

B-Phase Training Studies

In B-phase training studies, the target behavior is measured only during the treatment (B) phase. Because these designs lack a baseline (A phase), there is no systematic manipulation of the independent variable. Therefore the experimental effect cannot be demonstrated, thus abolishing any attempt at experimental control. It is therefore not possible to reliably determine the reasons for any change observed in the target behavior – it may be the result of treatment or it may not. One might reasonably ask the question as to why such an uncontrolled study that cannot demonstrate treatment effect is conducted in the first place. B-phase training studies are often implemented in situations of clinical emergency where the target behavior is at risk of causing harm to the patient or other people (Kazdin 2011). In this situation, there may not be sufficient time to conduct an initial baseline (A) phase and there may also be ethical objections to withdrawing the intervention (to revert to baseline to demonstrate treatment effect) if there has been amelioration of the target behavior.

Pre-intervention/Post-intervention Design

The pre-intervention/post-intervention design, wherein outcome measures are taken before and following (but not during) the intervention, appears to be an application of group-design methods to an individual. It does not conform to any of the tenets of single-case methodology and, as a research design for a single participant, is very weak. Backman et al. (1997) note that there is very little, if any, control for threats to internal validity in these designs. Although there is manipulation of the independent variable, the effects of such manipulation are not systematically recorded. That is, measures are not taken during the treatment (B) phase and consequently the experimental effect cannot be clearly demonstrated because changes in the outcome variable may be due to other confounds (e.g., practice effects, history, maturation) that have not been controlled. Moreover, in these types of studies, the outcome variable/s is usually measured with a standardized generic instrument rather than the specific and operationally-defined target behavior that is to be treated in therapy. There is no repeated and frequent assessment of outcome measures during all phases; rather assessments are conducted once (or on a small number of occasions) only prior to and after the conclusion of treatment.

Bi-phasic A-B Design

The bi-phasic A-B design can be considered the basic, entry-level design for single-case methodology. It is characterized by two phases: in the baseline (A) phase, the target behavior is measured repeatedly and frequently, and treatment does not commence until the B phase, in which the target behavior also continues to be measured repeatedly and frequently. Yet, because there is no second A phase (i.e., A-B-A), there is only one demonstration of the experimental effect and thus little, if any, control of extraneous variables. Without a control condition (e.g., a withdrawal or reversal phase, additional patients or behaviors that are concurrently examined) it is not possible to reliably establish the reason for change in the target behavior. Consequently, A-B designs are “more accurately considered a pre-experimental design” (Byiers et al. 2012, p. 401). Such designs cannot provide firm evidence to establish a causal (functional) relationship between the dependent and independent variables, even though it may be possible to verify statistically a change in level, trend or variability of the target behavior between the A and B phases.

Varieties of SCEDs in Psychology and Education with Examples from the Literature

Recent surveys have documented a rich diversity of designs reported in the current psychological and educational literature (Shadish and Sullivan 2011; Smith 2012; Barker et al. 2013). Barlow et al. (2009) describe 19 distinct SCEDs that allow investigations to be tailored to the diverse and often complex nature of behavioral interventions, and meet specific challenges in scientifically evaluating treatment effects. In this section, we review the four major types of SCEDs used in rehabilitation, behavioral and education research. The 19 design varieties described by Barlow et al. fit within the four major design types.

Withdrawal/Reversal Designs

Withdrawal/reversal designs constitute the 'basic' model of SCEDs. In general, they consist of a single sequence of alternating baseline (or no-intervention, A) phases and intervention (B) phases (see Fig. 3.2). Barlow et al. (2009) offer a comprehensive discussion of the advantages and limitations of the wide variety of withdrawal/reversal designs, including the more complex variants commonly reported in the literature, such as designs with additional A and B phases (e.g., A-B-A-B-A-B), or designs incorporating more than one intervention (e.g., A-B-A-C-A, where B and C denote different interventions).

Withdrawal/reversal designs are suitable for investigating interventions that are likely to have a reversible effect, so that the target behavior can be expected to return to baseline level when treatment is withdrawn. The most obvious examples of such interventions are equipment and external aids, such as communication devices, memory notebooks, personal digital assistants and so forth. A principal limitation of all withdrawal/reversal designs is that they cannot be used to examine interventions

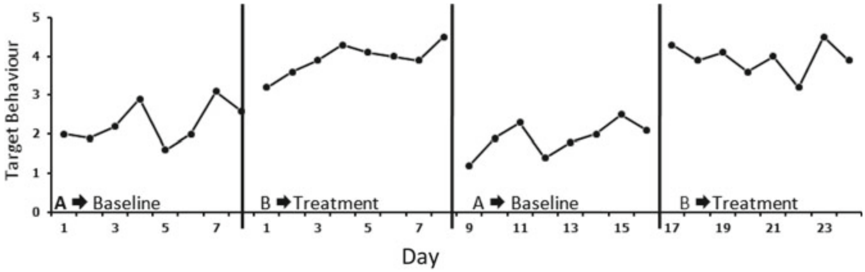


Fig. 3.2 Simulated data illustrating an A-B-A-B design

that have potentially irreversible effects on the target behavior, such as training the participant to use a self-monitoring strategy, which cannot be readily ‘unlearned’ when the treatment phase ends. These designs may also be unsuitable in situations when withdrawal of treatment might not be clinically appropriate or ethical, particularly if treatment appears to be effective or dangerous behaviors are involved. Moreover, if the investigation includes several interventions (e.g., A-B-A-C-A-D-A) the effect of intervention order, or interaction between different interventions may be difficult to interpret.

In general, N-of-1 trials in the medical literature are essentially multiple-phase, withdrawal/reversal designs, described by Guyatt and colleagues (1986, 1988) as the double-blind, randomized, multiple cross-over A-B trial in an individual patient. They may include washout periods following an intervention phase in order to reduce carryover effects of treatment. In addition, duration of treatment periods may be determined a priori in order to allow the expected treatment effects to occur (Duan et al. 2013).

Experimental effect in withdrawal/reversal designs is demonstrated if the level of the dependent variable (target behavior) varies in the predicted manner when the independent variable (intervention) is systematically introduced and withdrawn (i.e., at phase change from A to B, from B to A, etc.). Adequate experimental control is achieved when the study design provides at least three opportunities for demonstrating the experimental effect (Horner et al. 2005; Kratochwill et al. 2013). The A-B and A-B-A designs fail to meet this criterion. The A-B-A-B design (see Fig. 3.2) is now regarded as the simplest withdrawal/reversal design offering acceptable experimental control.

Travis and Sturmey (2010) used an A-B-A-B design to decrease the production of delusional utterances in a 26-year old man with “mild intellectual disabilities, frontal lobe syndrome, traumatic brain injury, mood disorder, and mania” (p. 745). Several years after sustaining his brain injury, this man began to utter delusional statements which negatively impacted the relationships he had with his peers in the inpatient facility where he lived. During baseline phases, the therapist provided 10 s of disapproving comment, immediately following the patient uttering a delusional statement. The intervention was based on differential reinforcement of alternative behavior and extinction. This consisted of withholding attention for 10 s when the patient uttered a delusional statement and providing 10 s of positive verbal reinforcement following contextually appropriate, non-delusional utterances. The behavioral experiment was conducted over 17 sessions with four to five sessions per phase, during which data were collected on the target behavior each session, and were presented graphically in the report. Compared with baseline performance, the intervention markedly decreased the rate (per minute) of delusional utterances, and increased the rate of non-delusional, contextually appropriate statements. Treatment effect was still evident at 6 month, 1- 2- and 4-year follow-ups.

Multiple-Baseline Designs

Many interventions teach new skills (e.g., communication strategies, gait retraining, social skills, behavior monitoring), which cannot be readily 'unlearned'. In this scenario, as well as the situation where it is considered unethical to withdraw a successful intervention, the multiple-baseline design (MBD) provides an effective and ready way by which to test the efficacy of an intervention. The MBD can also be used for interventions that can be meaningfully withdrawn.

In MBDs several baselines (legs or tiers) of the dependent variable are measured simultaneously. The intervention, however, is introduced across the various tiers in a staggered sequence. Thus, at different stages of the experiment some tiers will be in the baseline (A) phase and others will be in the intervention (B). There are three basic types of MBDs: across behaviors, participants or settings (see Fig. 3.3). Onset of the initial baseline phase occurs concurrently across all tiers, and onset of the first intervention phase is then sequentially staggered over time across the different tiers. Each tier generally consists of a simple A-B sequence, but more complex designs (e.g., alternating-treatment or multiphasic withdrawal/reversal designs) can also be

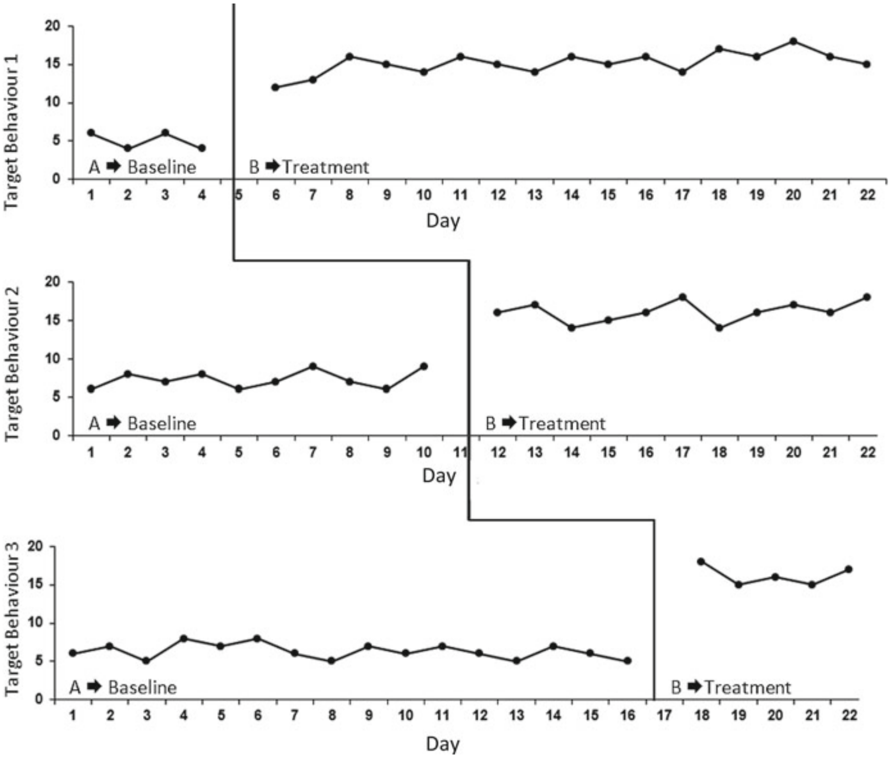


Fig. 3.3 Simulated data demonstrating a multiple-baseline design across three different settings

embedded into each tier (Shadish and Sullivan 2011; Smith 2012). In the latter instance, onset of additional phases is also generally staggered across the tiers.

MBDs designs are commonly reported in the psychology and education literature, accounting for more than 50 % of SCEDs (Shadish and Sullivan 2011). Experimental effect in MBDs is demonstrated when "*change occurs when, and only when*, the intervention is directed at the behavior, setting or participant in question" (Barlow et al. 2009, p. 202). Adequate experimental control is achieved when the design permits the experimental effect to be demonstrated on at least three occasions (Horner et al. 2005; Kratochwill et al. 2013). This means that in a MBD with three tiers, each tier must, at minimum, incorporate an A-B phase sequence.

The MBD has many strengths, efficiencies and flexibility in meeting specific contingencies of a situation which may cause problems with the classic withdrawal/reversal (A-B-A-B) design. They are most elegant in addressing replication (MBD across participants) and testing for generalization (MBD across settings and behaviors). Nonetheless, Kazdin (2011) has noted some potential difficulties with the MBD. There may be interdependence, particularly in MBD across behaviors, which means that when the intervention is introduced at the first tier, carry-over effects occur in the behaviors of the second and subsequent tiers which are still in the baseline phase. There may also be inconsistencies in response across tiers, which introduce ambiguity in interpreting results. In addition, prolonged baselines in the second and subsequent tiers may mean that in MBDs across behaviors or participants there is a lengthy period before intervention can commence.

Feeney (2010) used an A-B MBD across settings to investigate the effects of a multi-component intervention (including addressing environmental context, behavior supports and cognitive strategies) for reducing challenging behaviors in two children with traumatic brain injury. The intervention was delivered in three classroom settings: English Language Arts class, Mathematics class and Science class. In each setting, Feeney measured the treatment effect on three behavioral measures: (1) frequency of challenging behaviors (defined as attempted or completed acts of physical aggression such as hitting or pushing), or verbal aggression such as threats; (2) intensity of aggression measured on a 5-point scale (0 = no problems, to 4 = severe problems) using the sections relevant to aggression of the Aberrant Behavior Checklist; (3) percentage of work completed in the classroom. The experiment in both children lasted 30 days (the baseline in the first tier being 5 days), and data on the target behavior were collected for each of the 30 days and presented graphically. For both children, the treatment diminished both frequency and intensity of challenging behavior and also increased the quantity of work completed.

Alternating-Treatment Designs

In alternating-treatment designs (ATDs) the relative effect of two, or more, conditions is compared by administering each intervention to the same participant, over the same time span in an alternating sequence which can be quite rapid (e.g., within

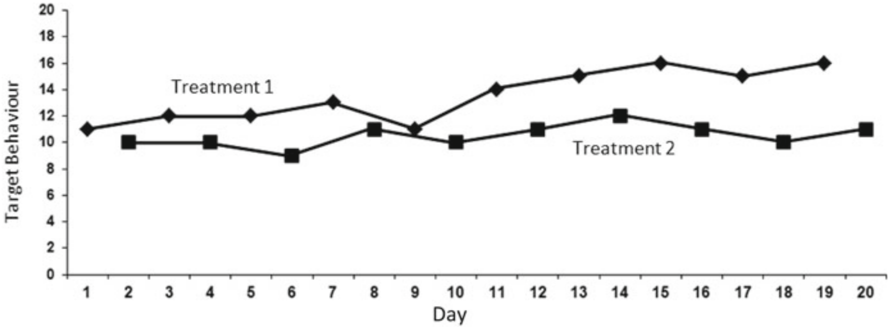


Fig. 3.4 Simulated data demonstrating an alternating treatment design, comparing two interventions. The treatment phase may be preceded by a baseline phase (e.g., Morrow and Fridriksson 2006), and/or a best intervention phase may follow the treatment phase

the same day) (Barlow et al. 2009). ATDs are particularly appropriate and useful when there is a need to identify an effective intervention as quickly as possible (see Fig. 3.4).

An initial baseline phase may precede the treatment phase (Morrow and Fridriksson 2006) and a final "best treatment" phase may also be included in the design. The "best treatment" phase refers to the final phase of the experiment where the intervention that has been shown to be superior is administered by itself. This phase is incorporated into the designs in order to evaluate threats to internal validity posed by potential interference associated with multiple interventions. Generally, factors such as treatment sequence, setting or therapists are also counterbalanced in order to reduce threats to internal validity. Experimental effect is demonstrated by "iterative manipulation of the independent variable (or levels of the independent variable) across observation periods" (Horner et al. 2005, p. 168). Experimental control is demonstrated when measures (levels) of the dependent variable for each intervention do not overlap.

ATDs are not suitable for investigating irreversible effects, and there is a risk that response generalization from one treatment to another may also occur. Interpretation of results may also be problematic due to carry-over effects or interaction between treatments. Intervention order may also influence outcome and make it difficult to attribute change in the dependent variable to a particular intervention.

Mechling (2006) used an alternating treatment design to compare the relative effectiveness of three interventions to teach three students (aged 5, 6 and 18 years) with profound physical and intellectual disabilities (assessed as functioning at levels between 6 and 13 months of age) to use switches to operate equipment. The interventions used different types of reinforces (intervention A = adapted toys and devices; intervention B = cause-and-effect commercial software; and intervention C = instructor-created video programs). The intervention was delivered over nine sessions, with sessions occurring 2–3 days per week. Each session lasted 9 min, within which the three interventions were administered, in block rotation order, for

3 min each. The dependent variable was the number of times the switches were activated. As is common with ATDs, there was no initial baseline phase in the study. Intervention C (instructor created video programs) was found to be the most effective intervention, and the study concluded with a "best treatment" phase, namely intervention C for three sessions.

Changing-Criterion Designs

Changing-criterion designs (CCDs) are a variant of MBDs. In CCDs, an initial baseline phase is followed by several intervention phases, each of which serves as a 'baseline' for the subsequent phase (see Fig. 3.5). CCDs are useful in the context of behavior shaping, when it is unlikely that the desired or expected magnitude of change in the target behavior (dependent variable) can be achieved using a one-step intervention. Performance criteria increasing in a step-wise manner are determined a priori, such that the magnitude of change between consecutive criteria is likely to be achievable. The content and structure of the intervention generally remain constant throughout the various intervention phases. Onset of a new phase occurs (i.e., the next performance criterion that becomes applicable) when the performance criterion for the current phase is reached. Experimental control is demonstrated by the repeated changes in the dependent variable as the criterion for performance is changed in a step-wise fashion at the completion of each treatment phase (Hartmann and Hall 1976).

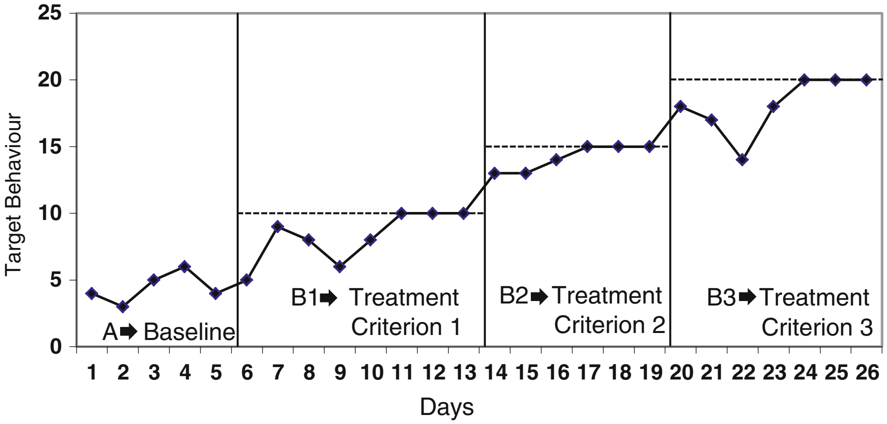


Fig. 3.5 Simulated data demonstrating a changing criterion design. *Vertical solid lines* indicate phase changes. *Dashed horizontal lines* indicate performance criterion level for a given treatment phase. Phase change occurs when criterion has been met, in this instance on three consecutive occasions

Skinner et al. (2000) used a CCD to evaluate the effectiveness of a shaping program aimed at improving leisure reading persistence in a young adult male diagnosed with paranoid schizophrenia. The dependent variable in the study was the number of pages read continuously by the participant. Following an initial baseline phase, the performance criterion for the first intervention phase was to read one page. The criterion was increased by one page at each phase change over the subsequent intervention phases, so that by the end of the treatment the patient was required to read eight pages. Onset of a new intervention phase occurred when the participant had reached the criterion on three occasions during the current phase. The reinforcer used in the intervention phases was chosen by the participant, namely earning a soft-drink when he read the required number of pages. Sessions (maximum one per day) were initiated at the participant's request throughout the study (he did so on 76 % of the 38 days he consented to participate in the study). A single, unplanned measure of maintenance obtained seven weeks post-intervention showed that the patient not only met, but surpassed the eight-page reading criterion.

Analysis of SCED Data

Compared with analysis of between-group designs data, agreement regarding data analysis in the context of SCEDs is less well established. As Smith (2012, p. 511) notes: "SCEDs undeniably present researchers with a complex array of methodological and research design challenges, such as establishing a representative baseline, managing the non-independence of sequential observations (i.e., autocorrelation, serial dependence), interpreting single-subject effect sizes, analyzing the short data streams seen in many applications, and appropriately addressing the matter of missing observations."

There have been two main approaches to analysis of SCED data: visual analysis and statistical analysis. Surveys of studies published in the behavioural sciences literature have shown that visual analysis was the predominant method of analysis during the 1960-1970s. By contrast, inferential statistics were used infrequently – between 4 % and 9 % of the articles sampled by Kratochwill and Brody (1978). More recent surveys (Barker et al. 2013; Smith 2012) have found a modest increase in the use of statistical techniques, although visual analysis remains the most common method. A comparable situation exists in the medical N-of-1 literature (Gabler et al. 2011). Some authorities suggest that statistical analysis using inferential statistics should be seen as complementary to, not a replacement for, visual analysis (Davis et al. 2013; Smith 2012). Randomization tests sit clearly within the realm of statistical analysis and are, in our view, a defining characteristic of the medical N-of-1 trial. For this reason, we discuss them separately from other statistical techniques.

Randomization Tests

Randomization tests are an important component of the statistical armamentarium that can be used to analyze data, and a fundamental aspect of experimental design. Yet, the term 'randomization tests' is somewhat of a misnomer. It does not refer to a specific statistical test or the calculation of a particular statistical parameter, but rather to the randomization of some aspect of the experimental design.

Briefly, there are three components to randomization in SCEDs. First, either treatment can be randomly allocated to time (i.e., to successive pairs of times or phases, to either half of the data collection period, or to specific blocks of time or phases), or the point in time at which treatment begins can be randomly selected (Edgington 1980). Random allocation in SCEDs helps to control for Type I errors and greatly enhances internal validity (Kratochwill and Levin 2010), and is considered by some to be the hallmark of *true* SCEDs (Barlow et al. 2009, p. 282).

Randomly allocating treatment to time lends itself well to pharmacological interventions. If allocation is also blinded (which is difficult to do in behavioral studies), the reliability is further strengthened. A short-coming of this strategy at least for behavioral studies, is that some of the possible combinations of treatment sequences may not be consistent with the study rationale. Randomly selecting the point at which treatment is to commence is perhaps more suitable in the rehabilitation setting but is also applicable in medical N-of-1 trials. Randomization of treatment commencement should be done such that the initial baseline contains at least eight observations (Edgington 1980) and subsequent phases comply with evidence standard requirements that all phases should contain at least three data points, (Barlow and Hersen 1984) and preferably five (Kratochwill et al 2013).

Second, a statistic is derived from the observed data. The choice of statistic is governed by the research question and the data limitations discussed below. For instance, if it is expected that treatment will improve the level of the dependent variable, then using a *t*-test to compare differences in mean score between phases may be appropriate.

Third, the selected statistic or index is calculated for all the possible permutations of the data generated during the randomization process. If nine possible permutations are generated by the randomization process (i.e., allocation of phase B commencement between, for example, observations 8 and 16 inclusive), the selected statistic is calculated for each of the nine possible data permutations The intervention is deemed to be effective if the probability is <0.05 that the statistic obtained for the randomly selected schedule (i.e., the one actually conducted) is greater (or lower, depending on the expected effect of the treatment) than the statistic obtained for the other possible schedules (see Wampold and Furlong 1981). At least 20 permutations ($1/20 = 0.05$) are needed before randomization provides sufficient power to detect differences that are significant at the 0.05 level.

A major limitation of randomization tests is that a larger number of observations are required in order to generate sufficient permutations and thus achieve adequate power than if conventional statistical tests were being used. Sufficient numbers of

Table 3.2 Standards of evidence for single-case experimental designs (After Kratochwill et al. 2010, 2013)

Features of graphed data to be examined within- and between-phases, evaluating the following:

1. Level
2. Trend
3. Variability
4. Immediacy of effect
5. Data overlap
6. Consistency of data patterns across similar phases

Procedure for visual analysis:	
Step 1:	Scrutinize the first A (baseline) phase to establish that the target behavior has been demonstrated to occur and there is an acceptable level of stability
Step 2:	Compare data in adjacent phases for level (phase mean), trend (slope of fitted line) and variability
Step 3:	Determine immediacy of intervention effect by examining changes in level, trend, variability, and degree of overlap between the last three data points in one phase and the first three data points in the next phase. Examine patterns of level, trend and variability for consistency across similar phases
Step 4:	Integrate information yielded in preceding steps to determine if there is adequate experimental control. If so, the intervention is deemed to work if data in either a treatment or baseline phase do not overlap the actual or extrapolated data pattern of the preceding baseline or treatment phase respectively

observations can be readily obtained in medical N-of-1 trials and random allocation is generally applied in this setting (Duan et al. 2013; Kravitz et al. 2014). By contrast, randomization is not, regrettably, commonly featured in behavioral SCEDs.

Visual Analysis

The traditional method of data analysis in SCEDs is visual analysis. It is fundamental to the meaningful analysis of SCED data and remains the most common approach used (Barker et al. 2013; Gabler et al. 2011; Smith 2012).

Visual analysis is defined as "reaching a judgment about the reliability and consistency of intervention effects by visually examining the graphed data" (Kazdin 1982; p. 232). Several guidelines for systematizing visual analysis have been proposed (e.g., Kratochwill et al. 2010; Lane and Gast 2013 – see Table 3.2), but there is still a dearth of agreed-upon operational decision-making criteria to guide the process.

Some proponents of this approach have argued that it should be the only method used because the effect of an intervention should be *clinically significant* and, hence, sufficiently clear in the graphed data so that statistical analysis will not be necessary to demonstrate that the intervention has worked. *Clinical* significance is concerned with whether or not an intervention has had a tangible (and supposedly beneficial) impact on the recipient and the way he/she functions and interacts in quotidian life (Kazdin 2001). An intervention effect might well be statistically significant, but so small that it has little or no impact on functional performance and be of little value to the recipient of the intervention.

As Kazdin (1978) points out, however, visual analysis has significant shortcomings: "The problem with visual inspection is that those individuals who peruse that data may not see eye to eye" (p. 638). First, there are no agreed upon formal criteria or decision rules. Inter-rater agreement, even among experienced judges, varies between not much above chance (De Prospero and Cohen 1978) to very high (e.g., Kahng et al. 2010), and is poorer for some features of the data, such as change in variability or slope, than others, such as change in level or mean (Gibson and Ottenbacher 1988). Using visual aids (e.g., celeration and trend lines) can increase the accuracy of visual judgments (e.g., Stocks and Williams 1995), but can also lead to misinterpretation by emphasizing the trend at the expense of other important features of the data (e.g., level; Brossart et al. 2006).

Visual analysis can also yield high rates of Type I error (up to 84 %), and the reliability of judgments can be confounded by variability in the data, autocorrelation[2] pre-existing linear trends, data cyclicity and effect size (Brossart et al. 2006; Jones et al. 1978). Although at least one study has reported high agreement (86 %) between statistical and visual analysis (Bobrovitz and Ottenbacher 1998), the two approaches are generally discordant, particularly when autocorrelation is high or when statistical analyses yield a significant result rather than a non-significant result (Jones et al. 1978).

Statistical Techniques

A variety of statistical techniques and approaches can be used to analyze SCED data, and these have been reviewed elsewhere (Brossart et al. 2006; Perdices and Tate 2009; Smith 2012). Techniques include the following: quasi-statistical techniques applied to graphed data (e.g., split-middle trend line; standard deviation band), randomization tests (described above), time-series analysis (e.g., C-statistic; autoregressive integrated moving average, ARIMA), traditional inferential statistics (e.g., parametric *t*-test; nonparametric Wilcoxon matched-pairs signed-ranks test; Friedman two-way analysis of variance), and effect sizes (using nonparametric methods, such as percentage of non-overlapping data; standardized mean differences; regression models; hybrid nonparametric/regression models, such as Tau – U; and multilevel modelling).

[2] Autocorrelation in a series of observations refers to the degree of predictability or lack of independence between one observation and the subsequent observation. It is usually expressed as a Pearson-Product Moment Correlation coefficient between all pairs of consecutive observations.

In contrast to visual analysis, statistical analyses follow a clear-cut set of operational rules and utilize replicable quantitative methods. Importantly, statistical procedures can provide a direct test of the null hypothesis (which visual analysis cannot) using a precisely defined significance criterion to determine if the treatment effect is reliable. Moreover, effect sizes can be calculated using statistical techniques, thus allowing findings from various studies to be synthesized (meta-analysis), whereas data pooling is not possible if visual analysis alone is performed. Kadzin (1982) suggests that statistical analysis should be used when the following occurs: (a) there is a notable trend or variability in the baseline, (b) a new treatment is being evaluated, (c) the treatment effect is not well understood, or (d) there is need for control of extraneous factors.

Yet, there is no “gold-standard” statistical analysis that is universally applicable to all and any SCED, nor agreement regarding the way in which effect sizes should be interpreted (Smith 2012). As is the case for between-group designs, the choice of statistical technique in SCEDs is largely dictated by the characteristics of the data. For example: (a) if there is a high degree of autocorrelation, parametric techniques are inappropriate because they assume independence of measures, (b) some techniques, such as the C-statistic, can have an unacceptable Type II error rate if there is a trend in the baseline and/or intervention phase data (Arnau and Bono 1998), (c) other techniques, such as the binomial test, yield poor control of Type I error (Crosbie 1993), (d) different methods are differentially sensitive to autocorrelation (Brossart et al. 2006), (e) techniques based on regression models might be more sensitive to changes in slope than changes in mean level across phases, whereas the reverse can be true for techniques based on General Linear models (Parker and Brossart 2003), (f) the size of the data set will preclude the use of some techniques, such as (ARIMA) which controls well for autocorrelation and confounders, but is only reliable when there are >50 observations (Box and Jenkins 1970), (g) in SCED data, where five or less observations per phase are not uncommon, violation of the assumption of normally distributed data is likely to be so great that parametric analyses are rendered inappropriate.

An important disadvantage of statistical analysis is that different techniques tend to produce different results. For example, the numerical magnitude of effect sizes will vary depending on the technique used to calculate them (Parker and Brossart 2003; Shadish et al. 2008). Moreover, there is no agreed-upon metric for interpreting effect sizes calculated for SCED data, nor is there consensus on how they relate to effect sizes calculated for between-group investigations (Kratochwill et al. 2013; Shadish et al. 2008; Smith 2012).

Conclusion

This chapter has described N-of-1 methods from the perspective of the behavioral sciences. In spite of its established history, however, the family of SCEDs has had a chequered course, both in the medical and behavioral sciences. We believe that this is contributed to, in part, by the proliferation of 'nonSCEDs' in the published literature (see Fig. 3.1), along with a poor understanding by many researchers and clinicians of the principles of the structure, design and implementation of single-case methods. Gladly, following the lead of Guyatt and colleagues (2002) more than 10 years ago, the medical randomized N-of-1 trial is now deemed Level 1 evidence for treatment decision purposes (Howick et al. 2011). At the same time, the behavioral sciences has also been busy in raising standards of design and evidence (Kratochwill et al. 2013), providing resources to plan, implement and critically appraise studies (Tate et al. 2013), and guidance to improve reporting (Tate et al. 2012). This recent confluence of events will assist in improving rigor in the conduct and reporting of SCEDs and augurs well for their future.

References

Allen KD, Friman PC, Sanger WG (1992) Small n research designs in reproductive toxicology. Reprod Toxicol 6:115–121

Arnau J, Bono R (1998) Short-time series analysis: C statistic vs Edgington model. Qual Quant 32:63–75

Backman CL, Harris SR (1999) Case studies, single-subject research, and n of 1 randomized trials: comparisons and contrasts. Am J Phys Med Rehabil 78:170–176

Backman CL, Harris SR, Chisholm J-AM, Monette AD (1997) Single-subject research in rehabilitation: a review of studies using AB, withdrawal, multiple baseline, and alternating treatments designs. Arch Phys Med Rehabil 78:1145–1153

Barker JB, Mellalieu SD, McCarthy PJ, Jones MV, Moran A (2013) A review of single-case research in sport psychology 1997–2012: research trends and future directions. J Appl Sport Psychol 25:4–32

Barlow DH, Hersen M (1984) Single case experimental designs. Strategies for studying behaviour change, 2nd edn. Allyn and Bacon, Boston

Barlow DH, Nock MK, Hersen M (2009) Single case experimental designs. Strategies for studying behaviour change, 3rd edn. Pearson, Boston

Bobrovitz CD, Ottenbacher KJ (1998) Comparison of visual inspection and statistical analysis of single-subject data in rehabilitation research. Am J Phys Med Rehabil 77(2):90–102

Box GEP, Jenkins GM (1970) Time-series analysis: forecasting and control. Cambridge University Press, New York

Brossart DF, Parker RI, Olson EA, Mahadevan L (2006) The relationship between visual analysis and five statistical analyses in a simple AB single-case research design. Behav Modif 30(5):531–563

Byiers BJ, Reichle J, Symons FJ (2012) Single-subject experimental design for evidence-based practice. Am J Speech Lang Pathol 21:397–414

Crosbie J (1993) Interrupted time-series analysis with brief single-subject data. J Consult Clin Psychol 61:966–974

Davidson PO, Costello CG (eds) (1978) N = 1: experimental studies of single cases. Van Nostrand Reinhold Company, New York

Davis DH, Gagné P, Fredrick LD, Alberto PA, Waugh RE, Haardörfer R (2013) Augmenting visual analysis in single-case research with hierarchical linear modeling. Behav Modif 37:62–89

De Prospero A, Cohen S (1978) Inconsistent visual analyses of intrasubject data. J Appl Behav Anal 12(4):573–579

Duan N, Kravitz RL, Schmid CH (2013) Single-patient (N-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol 66:S21–S28

Edgington ES (1980) Random assignments and statistical tests for one-subject experiments. Behav Assess 2:19–28

Feeney TJ (2010) Structured flexibility: the use of context-sensitive self-regulatory scripts to support young persons with acquired brain injury and behavioral difficulties. J Head Trauma Rehabil 25(6):416–425

Freud S, Breuer J (1895) Studies in hysteria (trans: Luckhurst N, Bowlby R). Penguin Books, London

Gabler NB, Duan N, Vohra S, Kravitz RL (2011) N-of-1 trials in the medical literature. A systematic review. Med Care 49(8):761–768

Gast DL (2010) Single subject research methodology in behavioural sciences. Routledge, New York

Gibson G, Ottenbacher K (1988) Characteristics influencing the visual analysis of single-subject data: an empirical analysis. J Appl Behav Sci 24:298–314. doi:10.1177/0021886388243007

Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S (1986) Determining optimal therapy—randomized trials in individual patients. N Engl J Med 314(14):889–892

Guyatt G, Sackett D, Adachi J, Roberts R, Chong J, Rosenbloom D, Keller J (1988) A clinician's guide for conducting randomized trials in individual patients. CMAJ 139:497–503

Guyatt G, Jaeschke R, McGinn T (2002) N-of-1 randomized controlled trials. In: Guyatt G, Rennie D, Meade MO, Cook DJ (eds) User's guide to the medical literature: a manual for evidence-based clinical practice, 2nd edn. McGraw Hill/AMA, New York/Chicago, pp 179–192

Hammond D, Gast DL (2010) Descriptive analysis of single subject research designs: 1983–2007. Educ Train Autism Dev Disabil 45(2):187–202

Hartmann DP, Hall RV (1976) The changing criterion design. J Appl Behav Anal 4:527–532

Horner RH, Carr EG, Halle J, McGee G, Odom S, Wolery M (2005) The use of single-subject research to identify evidence-based practice in special education. Except Child 71(2):165–179

Howick J, Chalmers I, Glasziou P, Greenhalgh T, Heneghan C, Liberati A, Moschetti I, Phillips B, Thornton H (2011) The 2011 Oxford CEBM evidence table (Introductory document). Oxford Centre for Evidence-Based Medicine. http://www.cebm.net/index.aspx?o=5653

Jones RR, Weinrott MR, Vaught RS (1978) Effects of serial dependency on the agreement between visual and statistical inference. J Appl Behav Anal 11(2):277–283

Kahng SW, Chung KM, Gutshall K, Pitts SC, Kao J, Girolami K (2010) Consistent visual analyses of intrasubject data. J Appl Behav Anal 43(1):35–45

Kazdin AE (1978) Methodological and interpretative problems of single-case experimental designs. J Consult Clin Psychol 46(4):629–642

Kazdin AE (1982) Single case research designs: methods for clinical and applied settings. Oxford, New York

Kazdin AE (2001) Almost clinically significant (p<.10): current measures may only approach clinical significance. Clin Psychol Sci Pract 8:455–462

Kazdin AE (2011) Single-case research designs: methods for clinical and applied settings, 2nd edn. Oxford University Press, New York

Kratochwill TR, Brody GH (1978) Single subject designs: a perspective on the controversy over employing statistical inference and implications for research and training in behavior modification. Behav Modif 2:291–307

Kratochwill TR, Levin JR (2010) Enhancing the scientific credibility of single-case intervention research: randomization to the rescue. Psychol Methods 15(2):124–144

Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, Shadish WR (2010) Single-case designs technical documentation. Retrieved from What Works Clearinghouse website. http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf

Kratochwill TR, Hitchcock J, Horner RH, Levin JR, Odom SL, Rindskopf DM, Shadish WR (2013) Single-case intervention research design standards. Remedial Spec Educ 34(1):26–38

Kravitz RL, Duan N, Niedzinski EJ, Hay MC, Subramanian SK, Weisner TS (2008) Whatever happened to N-of-1 trials? Insiders' perspectives and a look to the future. Milbank Q 86(4):533–555

Kravitz RL, Duan N, Vohra S, Li J (2014) Introduction to N-of-1 trials: indications and barriers. In: Kravitz RL, Duan N (eds) Design and implementation, Rockville of N-of-1 trials: a user's guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville

Lane JD, Gast DL (2013) Visual analysis in single case experimental design studies: brief review and guidelines. Neuropsychol Rehabil. doi:10.1080/09602011.2013.815636

Maggin DM, Chafouleas SM, Goddard KM, Johnson AH (2011) A systematic evaluation of token economies as a classroom management tool for students with challenging behaviour. J Sch Psychol 49:529–554

Mechling LC (2006) Comparison of the effects of three approaches on the frequency of stimulus activations, via a single switch, by students with profound intellectual disabilities. J Spec Educ 40:94–102

Morrow KL, Fridriksson J (2006) Comparing fixed- and randomized-interval spaced retrieval in anomia treatment. J Commun Disord 39:2–11

Moseley A, Sherrington C, Herbert R, Maher C (2000) The extent and quality of evidence in neurological physiotherapy: an analysis of the Physiotherapy Evidence Database (PEDro). Brain Impair 1(2):130–140

Parker RI, Brossart DE (2003) Evaluating single-case research data: a comparison of seven statistical methods. Behav Ther 34:189–211

Perdices M, Tate RL (2009) Single-subject designs as a tool for evidence-based clinical practice: are they are they unrecognised and undervalued? Neuropsychol Rehabil 19:904–927

Perdices M, Schultz R, Tate RL, McDonald S, Togher L, Savage S, Winders K (2006) The evidence base of neuropsychological rehabilitation in acquired brain impairment (ABI): how good is the research? Brain Impair 7(2):119–132

Rizvi SL, Nock MK (2008) Single-case experimental designs for the evaluation of treatments for self-injurious and suicidal behaviors. Suicide Life Threat Behav 38(5):498–510

Schlosser RW, Braun U (1994) Efficacy of AAC interventions: methodologic issues in evaluating behaviour change, generalization, and effects. Augment Altern Commun 10:207–223

Shadish WR, Sullivan KJ (2011) Characteristics of single-case designs used to assess intervention effects in 2008. Behav Res 43:971–980

Shadish WR, Rindskopf DM, Hedges LV (2008) The state of the science in the meta-analysis of single-case experimental designs. Evid Based Commun Assess Interv 3:188–196

Shamseer L, Sampson M, Bukutu C, Schmid CH, Nikles J, Tate R, Johnston BC, Zucker D, Shadish WR, Kravitz R, Guyatt G, Altman DG, Moher D, Vohra S, The CENT Group (2015) CONCORT extension for reporting N-of-1 Trials (CENT) 2015: explanation and elaboration. BMJ 350:h1793. doi:10.1136/bmj.h1793

Sidman M (1960) Tactics of scientific research. Evaluating experimental data in psychology. Basic Books, New York

Skinner CH, Skinner AI, Armstrong KJ (2000) Analysis of a client-staff-developed shaping program designed to enhance reading persistence in an adult diagnosed with schizophrenia. Psychiatr Rehabil J 24(1):52–57

Smith JD (2012) Single-case experimental designs: a systematic review of published research and current standards. Psychol 17(4):510–550

Stocks JT, Williams M (1995) Evaluation of single subject data using statistical hypothesis tests versus visual inspection of charts with and without celeration lines. J Soc Serv Res 20(3–4):105–126

Tate RL, McDonald S, Perdices M, Togher L, Schultz R, Savage S (2008) Rating the methodological quality of single-subject designs and N-of-1 trials: introducing the Single-Case Experimental Design (SCED) Scale. Neuropsychol Rehabil 18(4):385–401

Tate R, Togher L, Perdices M, McDonald S, Rosenkoetter U on behalf of the SCRIBE Steering Committee (2012) Developing reporting guidelines for single-case experimental designs: the

SCRIBE project. Paper presented at the 8th annual conference of the Special Interest Group in Neuropsychological rehabilitation of the World Federation of NeuroRehabilitation, Maastricht, July 2012. Abstract in Brain Impair 13(1):135

Tate RL, Perdices M, Rosenkoetter U, Wakim D, Godbee K, Togher L, McDonald S (2013) Revision of a method quality rating scale for single-case experimental designs and N-of-1 trials: the 15-item Risk of Bias in N-of-1 Trials (RoBiNT) Scale. Neuropsychol Rehabil 23(5):619–638

Tate RL, Perdices M, McDonald S, Togher L, Rosenkoetter U (2014) The design, conduct and report of single-case research: resources to improve the quality of the neurorehabilitation literature. Neuropsychol Rehabil 24(3–4):315–331

Travis R, Sturmey P (2010) Functional analysis and treatment of the delusional statements of a man with multiple disabilities: a four-year follow-up. J Appl Behav Anal 43(4):745–749

Vohra S, Shamseer L, Sampson M, Bukutu C, Schmid CH, Tate R, Nikles J, Zucker D, Kravitz R, Guyatt G, Altman DG, Moher D, The CENT Group (2015) CONCORT extension for reporting N-of-1 trials (CENT) 2015 statement. BMJ 350:h1738. doi:10.1136/bmj.h1738

Wampold BE, Furlong MJ (1981) Randomization tests in single-subject designs: illustrative examples. J Behav Assess 3(4):329–341

Chapter 4
N-of-1 Trials in Medical Contexts

Geoffrey Mitchell

Abstract Clinicians make treatment decisions on a regular basis, and some decisions may result in patients taking treatments for years. This decision-making is a core skill of clinicians, and if possible it should be evidence based. The most common tool to aid this decision making, the RCT, has many problems which can lead to a patient being prescribed a treatment that may not work for them. N-of-1 studies may be useful tools to assist in making the best decision possible. This chapter argues the case for N-of-1 studies assuming a place in the clinical armamentarium. It describes the rationale for and uses of N-of-1 trials, the advantages and limitations of N-of-1 trials, and discusses aggregation of N-of-1 trials to generate population estimates of effect.

Keywords N-of-1 trial • Randomized controlled trial • Clinical decision-making • Personalized care • Bias • Cross-over studies • Aggregated N-of 1 trial

Introduction

After a doctor has taken a history, examined the patient and possibly ordered and reviewed pathology tests and radiological examinations, a decision is made about the nature of the presenting problem. From this arises one of the most critical decisions to be made: how to manage the problem. This will often result in a medicine being prescribed. But, how does the doctor decide what treatment is best for that condition?

The following discussion relates to the likelihood of a treatment improving a presenting symptom, like pain or nausea. It does not relate to long-term treatments aimed at preventing a consequence like a stroke or heart attack. Here randomized controlled trials (RCTs) conducted on large populations and with long follow-up times are the only way of estimating benefit.

G. Mitchell (✉)
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: g.mitchell@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_4

How Clinicians Make Treatment Decisions

Therapeutic decision-making is not easy. Ideally the clinician will utilize published evidence for treatment efficacy, by either knowing the evidence for treatments, or searching for it amongst the vast academic literature at his or her disposal. Often there are clinical guidelines, which have been developed by expert reference groups who have identified and evaluated the literature and made considered recommendations. However, it is common that there is no credible evidence to guide a specific situation, and the clinician has to decide on relatively flimsy grounds. Sometimes clinicians choose a treatment on the basis of probable physiological or biochemical effect. While appearing logical, the reality may not match the theoretical effect. They may also take a "try it and see" approach, either on the basis of published trials, less robust evidence, or intuition.

Randomized Controlled Trials

Gold standard trial evidence comes from randomized controlled trials (RCTs). These are trials where the subjects are allocated a treatment purely by chance. The treatments in the trial are the test treatment, and either a comparator treatment in common use or a placebo, or dummy medicine. Sometimes the trial involves both groups being given the best available treatment, plus either the test treatment or a placebo.

In RCTs, subjects have an equal chance of being randomly assigned to the test treatment or the comparator. There are important reasons for testing a treatment in this way. Firstly, if a person is offered a treatment for a problem, they expect to observe an effect, whether or not an effect is actually there. This is called the placebo effect. In clinical practice, if a patient is prescribed a treatment and he or she experiences an improvement, it may be the placebo effect where taking a tablet leads to a presumed improvement. Alternatively, the observed improvement may have occurred, simply because the disease was resolving in line with its natural history. The illness may spontaneously resolve, as do upper respiratory illnesses caused by viruses.

Secondly, trials of treatment may demonstrate an improvement that is actually due to an unrelated factor. The patient may be taking an "over the counter" treatment, unknown to the doctor, and it might be impossible to tell which treatment, if any, was responsible for the resolution. Alternatively, treatment effectiveness may be influenced positively or negatively by some unrelated issue, termed a confounder, like age, gender or smoking status. If the sample size is large enough, randomizing the participants should lead to confounders being evenly distributed across the two groups, leading their effects on the trial outcome to be negated. The only thing that should influence the outcome is the test treatment.

Minimizing Bias

The whole purpose of RCTs is to try to eliminate the risk of the results being biased by something. There are a myriad of types of bias.

Some forms of bias (Higgins and Green 2011):

1. Selection bias. Participants are (consciously or unconsciously) allocated to one or other treatment group in the trial on the basis of the likelihood of improvement or poor likelihood of improvement, or some other parameters, like age, gender or appearance. This is prevented by a selection process that is truly random, like a computer generated randomization schedule. The randomization is done by someone completely at arms-length to the participants.
2. Performance bias. If those observing the patients are involved in their care, and know to which arm the person has been assigned, there is a risk that they may be wishing for a positive outcome in the trial, and (hopefully unconsciously) make observations in favor of one treatment over another. This is dealt with by blinding the allocation of the treatment to both the treating clinician and the person receiving the treatment, so called double blinding. This may not always be possible. The next best alternative is that the person doing the assessment of effect is not the treating clinician, but someone blind to the allocation.
3. Detection bias. This is where the results are derived in a manner in which they may be selectively reported. An example might be clinical records of blood pressure. The clinician (usually unconsciously) may record the blood pressure more often in a particular group of patients, perhaps on the assumption that it is more useful to do so. Such groups might include overweight or obese people compared with normal weight individuals, women on the oral contraceptive pill compared with those not on the pill, and so on.
4. Attrition bias. Here some people drop out of the trial in a differential way. For example, the trial treatment may give some people a headache, but the treatment is being tested for its ability to reduce their cholesterol levels. If the people who drop out are not taken into account in the analysis, a skewed result in favor of people who do not suffer headaches will occur. This is countered by so-called "intention to treat" analysis, where every person who enters the experiment is accounted for, and the characteristics of those who drop out are compared with those who stay. Further, a means of accounting for missing data in dropouts is devised so they are represented within the trial results.
5. Publication bias. Researchers are (unconsciously) less willing to report trials where the test treatment did not work, than those where the treatment was successful. Furthermore, journals are more likely to publish trials with positive results, than trials reporting no change in outcomes. Thus what is available to the clinician making the decision is not the full story. This is countered by the requirement to register the trials before commencement, so that interested people can use trials registries to track whether all or most of the trials being conducted are actually reported in the literature.

Problems Interpreting Trial Data

We have shown that there are problems in trial design, which researchers try very hard to minimize. However, decision-making is made even more difficult by the nature of clinical trials.

1. The evidence may be from clinical trials that have been conducted in different circumstances to that of the patient. Approved medicines have to have supporting evidence for a particular condition, a particular dose and a particular form of the treatment (for example, tablets, capsules or liquids). The most common issue is that the evidence is for a condition that is similar to, but not the same as, the patient's problem. In some cases the trial population is quite different to that of the patients. For example, a particular analgesic may have been tested in people with postoperative pain. It will be challenging to apply evidence derived from this setting to people with chronic pain seen in a physician's office.
2. Clinical trials often have very restrictive inclusion and exclusion criteria, so that the characteristics of the patients where benefit was displayed may be quite different to the characteristics of the patient in front of the doctor, even if the setting is the same as the trial. The classic example of this is where medicines with approval for use in adults are used in children. Another situation arises where the person's condition is similar to, but not the same as, the condition for which the approval has been obtained. Prescription in these circumstances is called off-label usage.
3. The evidence may be inferential, rather than actual trial evidence. This is virtually always the case in treatments for pregnancy. It is rare indeed for ethics approval to be granted for trials of a medicine to be conducted on pregnant women, because of the fear of adverse effects on the fetus. All that can be done is for studies to be conducted on pregnant animals, usually using doses far in excess of those to be applied in humans, looking for adverse effects on the fetus. For older medicines, clinical data on use and outcomes in pregnancy that may have been collected over many years can be used, but this is observational rather than trial based.

Randomized Controlled Trials (RCT) Report Population Averages

Trial data are based on population estimates of effect, but clinical decisions have to be made about individuals.

Arguably the most important constraint of RCT evidence is that it presents *population estimates of effect* – the mean effect of the treatment group is compared with that of the control group. However, *within* each of the groups, there will be a range of responses, and individuals may well have a contrary response to that of the bulk of people in that group. In Fig. 4.1, one of the intervention values falls within the confidence intervals of the control group and so the intervention actually had no effect. Three individuals in the intervention had such major effects that they fell outside of the upper 95 % confidence interval of the control group. These individuals responded so strongly to the comparator treatment (which might be placebo!) that they fell within the intervention confidence intervals – they had a very strong control treatment effect.

To try and overcome the problem of reporting average effects, different measures have been derived to help the clinician. These have in common an estimate of the *likelihood* that a treatment will work for a given person. Furthermore, because the control group did not have access to the intervention treatment, there is no way of knowing how they would respond if subjected to the intervention.

Cross-Over Studies

Cross-over studies involve the subject having both the intervention and treatment arms sequentially, and in random order. The main objective is to provide a single dataset from each participant in both the intervention and control arms. While it is possible to use the intervention and control arms to estimate the intervention response in the individual, this is usually not done. The data from a single pair could produce a result that does not reflect the actual participant response, simply by chance. Usually, trial data from a crossover study is not analyzed until the entire dataset is complete.

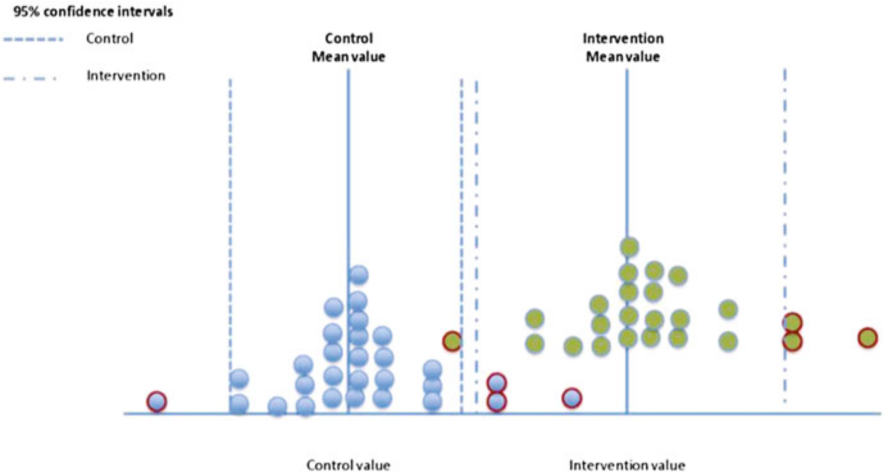


Fig. 4.1 The distribution of individual results in a randomized controlled trial (RCT). *Blue* values are control values, *green* are intervention values. Individual values that fall outside the confidence intervals of the other group have a *black border*

N-of-1 Studies

N-of-1 studies are double blind, placebo controlled multiple crossover trials measuring immediate treatment effects. They are described in detail in Sect. 1.1. The key differences compared with RCTs are that each participant receives both the intervention and comparator treatment in random order, and this is repeated multiple times – ideally at least three times. Each pair, termed a *cycle,* is analyzed. If the intervention treatment shows a stronger effect than the comparator in each cycle, this is the strongest evidence possible for the intervention in that participant. If the majority of cycles demonstrate a benefit for the test treatment, then the person is a probable responder. If more than one cycle favors the comparator, the patient is deemed a non-responder to the intervention treatment. This is one way of describing the results – there are others (See Chap. 9).

The trial design overcomes some key limitations of RCTs. In particular, N-of-1 studies allow all participants to receive both the active and the comparator treatments. Hence individual treatment decisions can be made with more certainty than those using RCT information.

Uses of N-of-1 Trials

Individual Decision-Making

For the reasons described above, RCTs have significant limitations. Most trial designs cannot estimate the efficacy of a treatment for an individual. There may be situations where this could be quite critical. These include: the treatment in question might be very expensive; there may be a significant side effect profile; or the treatment is controversial. It would be ideal to take such treatments only when a benefit will be obtained. N-of-1 studies have been used to assist in rational decision making for individual patients and their clinicians.

Example

A study has been completed which compares paracetamol with a non-steroidal anti-inflammatory medicine (NSAID), for chronic osteoarthritic pain of large joints (Yelland et al. 2006). NSAIDS can be very effective as treatment for osteoarthritis, but they carry a significant risk of gastro-intestinal bleeding, and of exacerbating both heart failure and renal impairment. Paracetamol has a more benign side effect profile, so may be an acceptable alternative so long as the clinical relief obtained is acceptable.

Each participant had three cycles, comprising 2 weeks of each of paracetamol and celecoxib in random order. To ensure blinding, every day they took active paracetamol, they also took placebo celecoxib, and vice versa. Every day they completed a symptom diary. At the end of the 12 week study period, the order within each pair was unmasked and the symptom diary data were analyzed. It was then possible to determine whether the patient's arthritis symptom improved with the NSAID or whether equal or better relief was obtained with paracetamol (Figs. 4.2 and 4.3).

What Treatments Have Been Tested Using Individual N-of-1 Studies?

Some of the treatments which have been tested in this way are found in Table 4.1. The technique has been used in symptom management in cancer treatment and palliative care, chronic non-malignant pain, Attention Deficit Hyperactivity Disorder, natural therapies vs prescription therapy for insomnia, and melatonin for sleep in children with ADHD.

The method can be used for any treatment where the following characteristics apply (Nikles et al. 2011):

- The treatment has to be expensive, or have a significant side effect profile, or is controversial (these trials can be complex to set up, and the effort has to be worth the cost);
- The condition is present at all times and has minimal fluctuations over time.
- The treatment does not alter the underlying pathology, but only treats symptoms;
- There is a short half-life;
- There is rapid onset of therapeutic effect and rapid reversal when the treatment is ceased;
- There is no cumulative treatment effect.

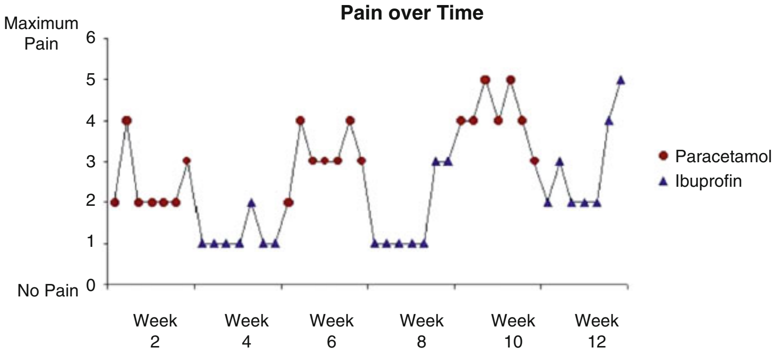


Fig. 4.2 Effect of non-steroidal anti-inflammatory medicines vs paracetamol for chronic arthritic pain in a responder to paracetamol

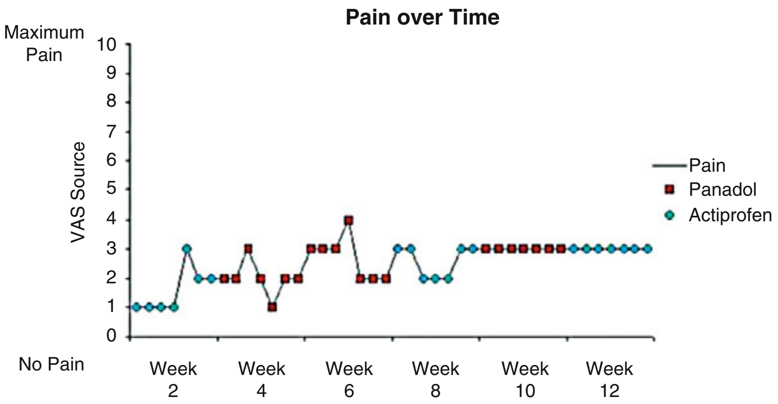


Fig. 4.3 Effect of non-steroidal anti-inflammatory medicines vs paracetamol for chronic arthritic pain in a non-responder to paracetamol

Table 4.1 Conditions and treatments where N-of-1 trials have been used by our group

Study	Principal author
Stimulant therapy for Attention Deficit Hyperactivity Disorder (ADHD)	Nikles et al. (2006)
Paracetamol vs celecoxib for chronic large joint arthritis	Yelland et al. (2007)
Paracetamol vs ibuprofen in chronic large joint arthritis	Nikles et al. (2007)
Temazepam vs Valerian for insomnia	Coxeter et al. (2003)
Gabapentin vs placebo in chronic neuropathic pain	Yelland et al. (2009)
Stimulants for fatigue in advanced cancer	Senior et al. (2013b); Mitchell et al. (2015)
Pilocarpine oral drops for dry mouth in palliative care	Nikles et al. (2013)
Paracetamol vs placebo in people already on opioids in palliative care	(Publication in preparation)
Stimulants for acquired ADHD in children with acquired brain injury	Nikles et al. (2014)
Melatonin vs placebo for children with ADHD	(In progress)

N-of-1 Studies Generating Population Estimates of Effect

While the initial intent of N-of-1 studies is to provide the strongest possible evidence for the effectiveness of a treatment in an individual, another use has emerged (Zucker et al. 1997; Nikles et al. 2011). This is to provide a population estimate comparable in power to a full RCT by aggregating a series of individual N-of-1 trials. The difference between the two trial designs is that in an RCT, the participant only receives either the active treatment or placebo/comparator. Potential participants may baulk at the prospect of receiving the comparator/placebo and withdraw or not participate. Since participants receive both test and comparator states, it they may be more willing to participate in aggregated N-of-1 trials if they find out whether the test treatment works for them.

Each participant in an N-of-1 trial contributes multiple datasets (often three or more) to each of the intervention and control arms of the trial Therefore, aggregating multiple N-of-1 studies is in effect a cluster randomized controlled trial, with the unit of the cluster being the individual patient. The number of participants required is far less than the equivalent RCT, and the two comparator groups are perfectly matched.

This technique can be used for some treatments where the treatment meets the characteristics described above. It may be an alternative to RCTs in patient groups where conducting a standard RCT is difficult or impossible (Nikles et al. 2011). For example, some populations are so small that there are not enough participants to generate the required sample size. This includes rare genetic conditions, rare cancers, and infrequent events like brain injury in children. A further scenario involves patients who are difficult to recruit or to retain in a trial, such as people approaching the end of life.

Example

Adults with advanced cancer frequently have fatigue, which is difficult to treat. The National Cancer Control Initiative of the USA has recommended stimulants (National Comprehensive Cancer 2010), but the evidence base is scant. We undertook an aggregated N-of-1 study of methylphenidate (MPH), which yielded 43 patients and 84 completed cycles. The equivalent RCT would have required 94 participants. The population estimate showed that there was no difference between treatment and placebo. However, because each participant contributed data to both intervention and control data, a report for each patient was generated. We showed that, although there was a negative population finding, seven patients had important differences favoring MPH over placebo, and one patient had important worsening of fatigue on MPH (Senior et al. 2013a; Mitchell et al. (2015) (Fig. 4.4).

Clinical Advantages of N-of-1 Trial Design

N-of-1 trials produce results that can guide individuals and their clinicians to make rational treatment choices. While this is easiest when all three cycles show results in favor of the intervention, it is also useful when two of the three interventions favor the intervention, but the level of uncertainty rises because of the risk that the results could have arisen by chance (Nikles et al. 2000).

These trials can be complex to establish and run. However, the advantage to the patient is more certainty that the treatment will work for them. The patient also has the advantage of *not taking* medicines that are known *not* to work, which reduces patient costs and the risk of adverse events and drug interactions with the test medicine.

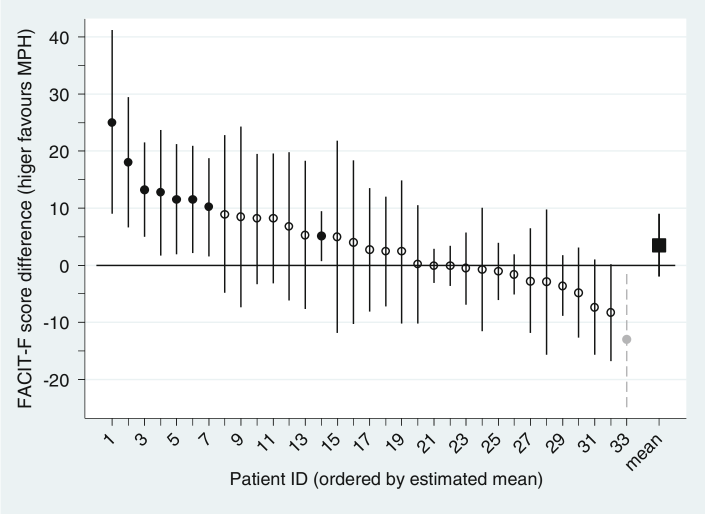


Fig. 4.4 Mean difference (95 % credible intervals) between methylphenidate (MPH) compared to placebo on individual fatigue scores (FACIT-F) for each patient (*circle*) and the overall group (*square*) (Note: the *solid black circles* designate positive responders, the *hollow circles* designate non-responders, and the *solid gray circle* designates a negative responder)

N-of-1 trials represent a major shift in the way clinicians could think about treatment decisions. Given the time pressures that most work under, rapid decision-making tends to override the benefits of certainty that treatments can work by conducting an N-of-1 trial. However, treatment decisions can have longstanding consequences. The decision to commence stimulant therapy for a child with ADHD for example, could lead to years of therapy. To have at the clinician's disposal a means of determining if the treatment is effective or not before the decision to prescribe long-term is made, should lead the clinician to make use of this opportunity. This has been borne out in a study of treatment decisions in children with ADHD, who underwent N-of-1 trials of stimulant therapy (Nikles et al. 2006). Forty-five doctors across Australia requested 108 N-of-1 trials, of which 86 were completed. In 69 drug-versus-placebo comparisons, 29 children responded better to stimulant than placebo. Immediately post-trial, 19 of 25 (76 %) drug-versus- placebo responders stayed on the same stimulant, and 13 of 24 (54.2 %) non-responders ceased or switched stimulants. In 40 of 63 (63.5 %) for which data were available, post-trial management was consistent with the trial results. For all types of N-of-1 trials, management changed for 28 of 64 (43.8 %) children for whom information was available. 12 months after the trial, 89 % of participants were still adhering to the treatment consistent with the trial, with the concordance rate falling from 50 % at the time of trial to 38 % (Nikles et al. 2007) (Figs. 4.5 and 4.6).

Another use of N-of-1 studies is to assist in determining if an existing treatment is working or not. Professor Dave Sackett, one of the founding fathers of evidence based practice, established a single patient trial clinic at his hospital. Its purpose was

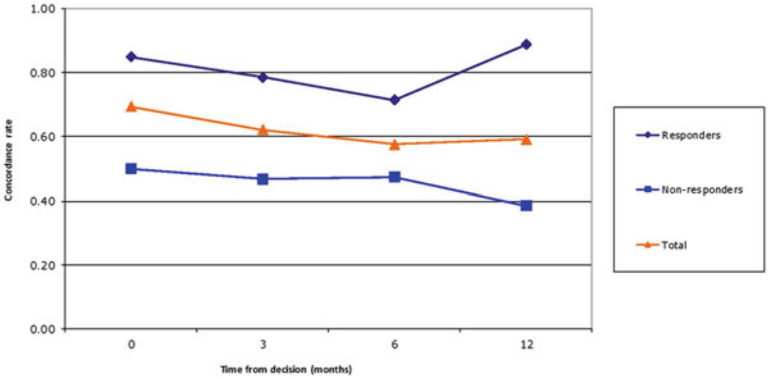


Fig. 4.5 Stimulant vs comparator trials in ADHD children – concordance rate from time of N-of-1 test result

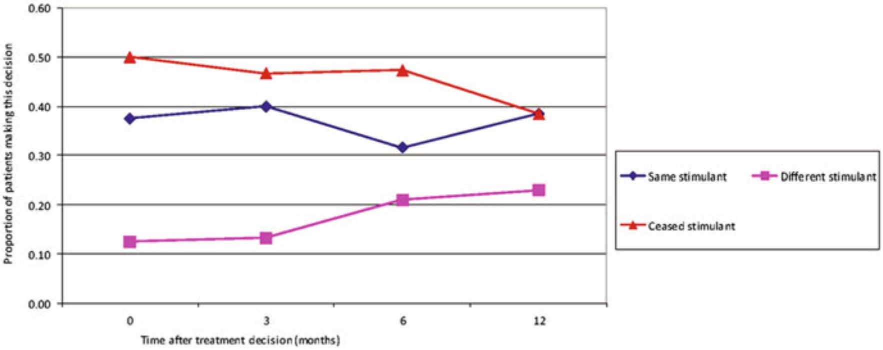


Fig. 4.6 Treatment decisions by time after N-of-1 test results received – non-responders

to utilize the method to help solve challenging clinical dilemmas. This clinic arose from a case of intractable asthma where the clinicians thought there was a better response to one treatment (theophylline) than an alternative one (ipratropium). Sackett, using an N-of-1 approach, showed that the patient felt *worse* on theophylline than when not taking it (Sackett 2011). The opportunities that such clinics could create in terms of higher quality treatment, better decision making, improved patient outcomes, and reduced system costs are obvious.

Limitations of N-of-1 Trials

N-of-1 studies are not a panacea. They are useful only when certain conditions are met in the treatment to be tested, as discussed above. They can be complex, and it may take considerable setting up to ensure they are done well. If there is access to a service that can set them up for the clinician, this could make it simpler – more like ordering a pathology or radiology test.

The use of evidence derived from aggregating N-of-1 studies is limited because few of these tests have been done. In addition, there is overwhelming acceptance of RCTs as the gold standard, and the place of neither individual nor aggregated N-of-1 studies is not clear in the minds of most clinicians. They are yet to find their place in the clinical armamentarium. Finally, there is a perceived risk that the results of aggregated N-of-1 studies may not be generalizable. The very characteristic that makes them attractive in situations where gathering evidence is difficult – small participant numbers needed to get adequate statistical power – may lead to the sample not being representative of the broader population in question. In fact, this is the case for all RCTs, particularly of symptom interventions. It is important if possible to compare the test population with the characteristics of the broader population in question, to try and ensure that the results are generalizable.

Conclusion

Clinicians make treatment decisions on a regular basis, and some decisions may result in patients taking treatments for years. This decision-making is a core skill of clinicians, and if possible it should be evidence based. The problem is that the most common tool to aid this decision making, the RCT, has many problems which can lead to a patient being prescribed a treatment that may not work for them. N-of-1 studies may be useful tools to assist in making the best decision possible. This chapter argues the case for N-of-1 studies assuming a place in the clinical armamentarium. Future chapters will look in detail at how this can be done.

References

Coxeter PD, Schluter PJ, Eastwood HL, Nikles CJ, Glasziou PP (2003) 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 11:215–222

Higgins J, Green SE (2011) Cochrane handbook for systematic reviews of interventions. Version 5.1.0. The Cohrane Collaboration, London

Mitchell G, Hardy J, Nikles J, Carmont S, Senior H, Schluter P, Good P, Currow D (2015) The effect of methylphenidate on fatigue in advanced cancer: an aggregated n-of-1 trial. J Pain Symptom Manage. doi:10.1016/jpainsymman.2015.03.009

Network NCC (2010) NCCN clinical practice guidelines in oncology- adult cancer pain V.1.2010. NCCN, Washington, DC

Nikles CJ, Glasziou PP, Del Mar CB, Duggan CM, Mitchell G (2000) N of 1 trials. Practical tools for medication management. Aust Fam Physician 29:1108–1112

Nikles CJ, Mitchell GK, Del Mar CB, Clavarino A, Mcnairn N (2006) An n-of-1 trial service in clinical practice: testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. Pediatrics 117:2040–2046

Nikles CJ, Mitchell GK, Del Mar CB, Mcnairn N, Clavarino A (2007) Long-term changes in management following n-of-1 trials of stimulants in attention-deficit/hyperactivity disorder. Eur J Clin Pharmacol 63:985–989

Nikles J, Mitchell GK, Schluter P, Good P, Hardy J, Rowett D, Shelby-James T, Vohra S, Currow D (2011) Aggregating single patient (n-of-1) trials in populations where recruitment and retention was difficult: the case of palliative care. J Clin Epidemiol 64:471–480

Nikles J, Mitchell GK, Hardy J, Agar M, Senior H, Carmont SA, Schluter PJ, Good P, Vora R, Currow D (2013) Do pilocarpine drops help dry mouth in palliative care patients: a protocol for an aggregated series of n-of-1 trials. BMC Palliat Care 12:39

Nikles CJ, Mckinlay L, Mitchell GK, Carmont SA, Senior HE, Waugh MC, Epps A, Schluter PJ, Lloyd OT (2014) Aggregated n-of-1 trials of central nervous system stimulants versus placebo for paediatric traumatic brain injury–a pilot study. Trials 15:54

Sackett DL (2011) Clinician-trialist rounds: 4. why not do an N-of-1 RCT? Clin Trials 8:350–352

Senior HE, Mckinlay L, Nikles J, Schluter PJ, Carmont SA, Waugh MC, Epps A, Lloyd O, Mitchell GK (2013a) Central nervous system stimulants for secondary attention deficit-hyperactivity disorder after paediatric traumatic brain injury: a rationale and protocol for single patient (n-of-1) multiple cross-over trials. BMC Pediatr 13:89

Senior HE, Mitchell GK, Nikles J, Carmont SA, Schluter PJ, Currow DC, Vora R, Yelland MJ, Agar M, Good PD, Hardy JR (2013b) Using aggregated single patient (N-of-1) trials to determine the effectiveness of psychostimulants to reduce fatigue in advanced cancer patients: a rationale and protocol. BMC Palliat Care 12:17

Yelland MJ, Nikles CJ, Mcnairn N, Del Mar CB, Schluter PJ, Brown RM (2006) Celecoxib compared with sustained-release paracetamol for osteoarthritis: a series of n-of-1 trials. Rheumatology (Oxford) 46:135–140

Yelland MJ, Nikles CJ, Mcnairn N, Del Mar CB, Schluter PJ, Brown RM (2007) Celecoxib compared with sustained-release paracetamol for osteoarthritis: a series of n-of-1 trials. Rheumatology (Oxford) 46:135–140

Yelland MJ, Poulos CJ, Pillans PI, Bashford GM, Nikles CJ, Sturtevant JM, Vine N, Del Mar CB, Schluter PJ, Tan M, Chan J, MacKenzie F, Brown R (2009) N-of-1 randomized trials to assess the efficacy of gabapentin for chronic neuropathic pain. Pain Med 10:754–761

Zucker DR, Schmid CH, McIntosh MW, D'Agostino RB, Selker HP, Lau J (1997) Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. J Clin Epidemiol 50:401–410

Chapter 5
Aggregated N-of-1 Trials

Geoffrey Mitchell

Abstract The original purpose of N-of-1 trials is to determine whether a treatment works in a person. However, these trials can be considered as mini-randomized controlled trials (RCTs), with the person providing multiple datasets to the intervention and control groups. Therefore, several people undergoing the same N-of-1 trial can contribute many data sets and this rapidly scales up to the point where the power of the trial can equate to a normal RCT, but with far fewer participants. This characteristic means that RCT-level evidence can be derived from populations that are almost impossible to gather data from, because of low prevalence conditions, or difficulty in recruiting or retaining subjects. This chapter describes the method in detail, along with methodological challenges and limitations of the method.

Keywords N-of-1 trials • Randomized controlled trials • Methodology sample size • Error types • Confidence intervals • Aggregated N-of-1 trial • Representativeness • Adverse events • Selection bias

Clinical Care and Evidence

The catch-cry of modern healthcare is that treatments applied should be based on evidence of effectiveness. Treatments should be administered only if they have been tested and proved not to be harmful, and provide a benefit. The gold standard treatment in this regard is the randomized controlled trial (RCT). The objective is to test a treatment in such a way that the only thing that influences the result is the test treatment itself, and conditions that may modify the result are neutralized as much as possible. These include biases towards or against the treatment, and conditions that may confound the result by producing a similar effect as the test treatment, or impacting on the treatment outcome in other ways.

Participants in an RCT are randomly allocated to one or other of the treatments or comparators. The results for each group are then analyzed and group differences

G. Mitchell (✉)
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: g.mitchell@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_5

are reported. Clinicians use these results to decide whether to use the treatment or not, based on these population results.

Please note that the following discussion is only relevant to evidence about treatment interventions, and not large scale, long term, population-based outcomes. Where the outcome is death or a major morbid event, thousand of people have to be followed for years to identify a difference in mortality or morbidity between intervention and comparator treatments.

The Importance of Sample Size

A critical element in determining the success or otherwise of these trials is whether a pre-determined sample size is achieved. The sample size is a participant number required in order to minimize the risk of random error.

There are two types of error (Bland 1995). The first (Type 1) is saying a treatment does work when in fact it does not. This is addressed by recognizing that drawing a random sample from a pool of like participants may by accident assemble a group who do not respond to the test treatment as the population as a whole would. Statistical measures are used to estimate the likelihood that the result seen is actually the effect of the treatment. Common descriptions of this are the p value and 95 % confidence intervals. The p value is in effect the likelihood that this is not a true result. Hence the statement $p<0.05$, says that the likelihood that this result is not representative of the true effect, is less than 5 %. The smaller the p value, the lower is the probability that the result has arisen by chance. The term 95 % confidence interval looks similar but is in fact the opposite to p values. It is a positive statement of the likelihood that this is a true result. Say a blood pressure treatment is tested, and the result is stated as a reduction in systolic blood pressure of 7 mm mercury (95 % confidence intervals (or CI) 4, 10). This means that if the result had fallen anywhere between 4 and 10 mm mercury, the reader can be 95 % certain that this is a true result.

Type one errors are serious because treatments are being applied with the intent of improving a situation, and if they are being applied on the basis of trial evidence of effectiveness and do not work, a lot of resources are wasted. More critically, people may be exposed to the risk of adverse effects of the treatment for no benefit.

The second error (Type 2) is to say that a treatment did not work when in fact it did. Here there may be a small difference in effect, which did not reach significance (the measures of effect did not reach the pre-determined definitions of effectiveness). The way this error is avoided, is to determine how large the sample should be in order to detect a predetermined effect size with a predetermined level of confidence that the result is true. (This is termed the power of the study). If the trial does not reach that sample size, and the result is not significant, the reader cannot tell if this because there was really no effect, or that there was an effect and the trial was too small to detect it. Therefore the success or failure of a trial can be impacted by the success or failure of the trial to reach the predetermined sample size.

Sample size is determined by the size of the probable effect of the intervention. If there is an expected large effect size, then the sample size is small. Conversely the sample size may be in the hundreds or thousands if the effect or event sought has a low prevalence. This is the case for outcomes like deaths or life threatening events where to a particular treatment is trying to avoid these. An example might be a treatment to prevent deep venous thrombosis during airline flights. Finally, if the impact of a limited intervention on a parameter such as quality of life is likely to be small, a large sample will be needed to demonstrate it.

Difficulties in Attaining the Predetermined Sample Size

It can be very difficult to reach the predetermined sample size. There may be stringent inclusion and exclusion criteria, which make most potential participants ineligible. It may be necessary to screen many people to find one eligible participant. Trials may require long recruitment times and very large budgets to achieve the sample size, even if the condition is relatively common.

There are some situations where achieving the sample size is exceedingly difficult (Nikles et al. 2011). For example, the condition may have a *low prevalence*. Say a trial is planned where the estimated sample size for an RCT is 250. If there are only 500 people in the country with the condition, the recruiters would have to recruit half of all eligible people into the trial. This is virtually impossible.

There may be conditions where it is very *difficult to recruit or retain people*. Palliative care is the classic example. The subjects are very sick and can deteriorate very quickly. Hence relatively small numbers may agree to participate (Davis and Mitchell 2012). Even if they agree, they may not stay well enough to complete the trial, and their data are lost when they withdraw or die. Gate-keepers may limit access to potential participants on the basis that the patient is too sick. Formal RCTs in palliative care populations often require multiple recruitment sites, significant staffing and long recruitment time frames to achieve the required sample size (Shelby-James et al. 2012).

Aggregating N-of-1 Trials

Individual N-of-1 trials have been described elsewhere in the book (Chaps. 3 and 4). The characteristics that are so useful in providing information about the efficacy of a treatment in the individual can be used to provide population estimates of effect (Nikles et al. 2011). An individual N-of-1 study could be considered as a series of RCTs in an individual. Each cycle is a double blind, placebo- or comparator-controlled trial, repeated multiple times.

Therefore, an N-of-1 trial comprising three cycles could be considered as a series of three RCTs, where the participants are perfectly matched. If multiple people do

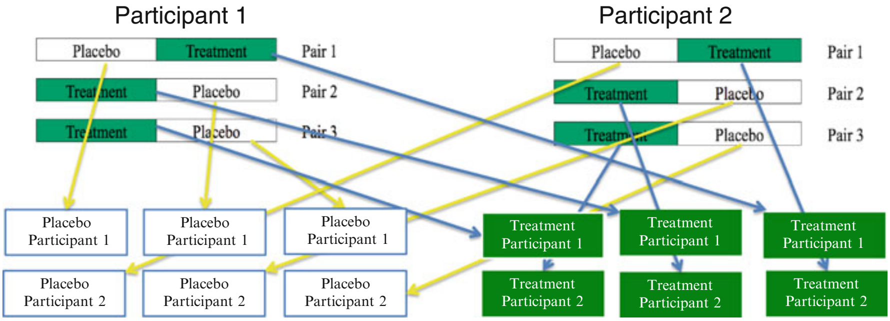


Fig. 5.1 Aggregated N-of-1 trials contributing multiple datasets to a virtual randomized controlled trial. Effectively, each participant provides multiple datasets to each side of an RCT

the same trial, then it can be considered as a cluster RCT, with the unit of clustering being the individual participant.

Hence if each person participating does a three cycle N-of-1 trial, they contribute three sets of data about the test treatment, and three for the comparator (Fig. 5.1). Even accounting for a cluster effect, there is a dramatic escalation in the accumulation of trial data with the addition of each new person to such a trial.

Advantages of Aggregated N-of-1 Studies Over RCTs

Sample Size

There is a real opportunity to generate very high quality evidence in populations that are hard to recruit to or to retain, with a much smaller sample size. The importance of this is hard to overstate. In Australia it has only recently been possible to conduct fully powered RCTs in palliative care, due to the development of a national multisite clinical trials network in Australia. This has only been possible through substantial government funding. Important evidence is being gathered (Hardy et al. 2012). For many treatments in palliative care, the rationale for using treatments is based on evidence in related but more high prevalence populations (e.g. cancer patients receiving chemotherapy), by inference using knowledge about physiological mechanisms (e.g. which anti-emetic to use), or guesswork. If credible evidence can be generated with small populations, it will improve treatment quality in this area. The same can be said for other populations, For example, an N-of-1 trial of stimulants in traumatic brain injured children required 50 children providing between 118 and 128 completed cycles. The same parallel group RCT required 242 participants.

The Arms of the Treatment and Comparator Groups are Perfectly Matched

The reporting of RCTs always includes description of the characteristics of the groups randomized to intervention and control. If most parameters (e.g. proportion of females, mean age, smoking status) are similar as evidenced by a non-significant difference between the trial groups, it is considered proof that randomization has worked. The likelihood that recruitment bias has occurred is considered low.

In aggregated N-of-1 trials, because each individual contributes data to both the intervention and comparator sides of the trial, the groups are perfectly matched. This allows for a smaller sample size to be required, and removes the possibility of participant bias.

Usefulness of Data When Participants Withdraw

If a participant drops out part way through an RCT, or is lost to follow-up, those data are effectively lost to the trial. This is an expected problem, and the sample size is inflated to account for this. If a patient does drop out, but completes at least one cycle of data in an N-of-1 trial, then the completed cycles can be added to the final dataset for analysis. The sample size can be described in terms of the number of participants required, *and* the number of cycles completed. This reduces the number of patients required to be recruited, and shortens the trial and reduces the cost.

Participant Receives a Result Immediately

Conducting multiple comparisons of test and comparator treatments for each participant means that shortly after a participant's trial has finished, a report can be generated about the effectiveness of that treatment in that patient. Unlike RCTs, where pooled data are available after the trial has closed, each participant gets an immediate benefit from having participated in the trial. He or she and their clinician can make an informed treatment choice. Another advantage of being exposed to both arms of a trial, is that patients do not have the disincentive that they have a chance to be exposed to the control arm of the trial only.

The Results Describe What Happens to Everyone Within a Trial

The result of a parallel arm RCT is the mean effect and standard deviation of each group, and a decision is made about effectiveness based on the difference in effect between the groups. No inference can be made about what happens to individuals

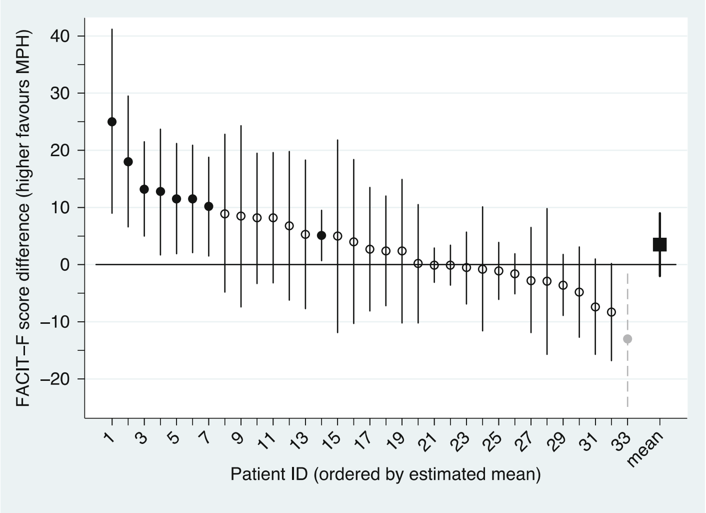


Fig. 5.2 Multiple N-of-1 trials of methylphenidate vs placebo for cancer fatigue. Participants with shaded dots show clear improvement, most show no change (*unshaded dots*) and one (*grey dotted lines*) is worse. All this is hidden within a single population estimate

because any one individual only gets exposed to one arm of the trial. Contrast this with the results seen in the aggregated N-of-1 trial. Figure 5.2 shows the results of an N-of-1 trial of methylphenidate in fatigue in palliative care patients (Mitchell et al. 2015). The mean and confidence intervals of the group are small. However, note the variation in individual responses that are in effect hidden within this population result.

N-of-1 studies allow an indication of what happens within a trial. How many will fall outside the result as described by the population result, and actually benefit from the test treatment? There will also be some who get worse on the treatment.

It could be argued that this is a major weakness of parallel arm RCT results where the trial is testing an intervention for a symptom. There may be a case for developing a technique where an estimate of the proportion of people who will respond is made. This could be done by adding a crossover element to the trial, for suitable treatments. The original calculations would be done on the first crossover, and the population estimate calculated. The proportions that appeared to respond and get worse could be estimated. The trial would report both the population response and the estimated proportions who appeared to respond or get worse. Because there is only one crossover, the level of precision is not as high as in a multiple crossover design. This suggestion is a pragmatic balance between precision and the practicalities of lengthening a trial with the added expense and time this would entail.

Then the discussion between doctor and patient around treatment decisions could be couched as follows for a trial with a null treatment effect: "Most people in the population will not derive a benefit. However, it is estimated that (say) 20 % will benefit, and (say) 5 % will get worse if they try it. Would you consider a trial of this treatment where you have those odds of a positive or a negative response?"

In Chap. 16 we discuss how to analyze aggregated N-of-1 trials.

Problems With Aggregated N-of-1 Trials Compared With RCTs

The Trial Population May Not be Representative

While a small sample size allows statistically credible evidence to be gathered, it also raises the prospect that the result will not be viewed as credible. The first objection is that the participant numbers may be unrepresentative of the broader population it represents. This is a fair comment, but the same can also be said of a larger RCT as well. All that can be done is to describe the cohort as fully as possible. All RCTs define their population with the view of showing equivalence between intervention and comparator groups. In a similar way, the researchers should attempt to compare the demographics of those selected to participate with the demographics of larger cohorts of the same population.

Inability to Determine Potential Predictive Factors for Responders and Non-responders

Assume a trial determines that some people respond to a treatment and not others. Two distinct groups therefore exist in the trial population, and it should be possible to compare demographic and clinical characteristics of the two groups, looking for markers likely to predict response.

However, the study is usually powered for the primary outcome – a change in a clinical condition like pain. Unless it was powered to identify characteristics, there is a risk of Type 2 error when reporting characteristics of responders. If there are statistically significant characteristics present in responders with these smaller numbers, then they are truly representative of clinical differences between the intervention and control group. Then the above argument arises: do these observed differences apply to a larger population?

The very small participant number runs a similar risk to an underpowered study in some people's minds– does a negative result mean it is not possible to say that it truly represents the outcomes for a person to whom the study results could be applied, and would a larger sample give a more reliable result?

Inability to Detect Low Prevalence Harmful Side-effects

The same sample size problem applies to this issue. Is the sample size big enough to detect low prevalence events, as many adverse effects of treatments are? It may not be and therefore there is a risk of reporting no adverse events, when the risk of a potentially dangerous adverse event may exist. There is no counter to this problem, except to demonstrate a robust process to record and assess any adverse events that do occur. Minor adverse events can be recorded, and serious adverse events need to be reviewed by an independent drug safety monitoring committee (See Chap. 10). All that can be done is to state in the discussion that serious side effects may exist, but were not observed in the test population.

Risk of Selection Bias

It is possible that by presenting individual results to the patient and clinician, that the clinician may start to think that certain patients will respond to the treatment and others will not. They may then present certain patients for recruitment and avoid others. For this reason it is ideal to have one person recruiting people and a different person presenting the results to the patient and another seeing the completed data and making treatment decisions. Obviously this may not be practical. Tests of whether there are trends in recruiting responders and non-responders over the course of the trial, to detect whether the proportions of each change over the course of the trial, have to be conducted if the one clinician is both selecting patients and discussing the trial results.

For the same reason it is important that the person generating individual reports is not the person analyzing the completed dataset. This is particularly the case if the technique of determining an important clinical effect is based on the posterior probability of effect. The posterior probability figure is the one used to decide whether the person has responded to a treatment or not, and will constantly change as new patient data are added. This requires constant analysis of the data by the researcher. Whether or not the participant is considered a responder to the treatment has major implications for the patient. It is possible that the reports presented to the patient and clinician could be influenced by the knowledge of what has been observed in the posterior probability of effect.

Planning the Trial

Determining the Sample Size

This is dealt with in Chap. 16.

As with other trials, the most important information required is an estimate of the expected difference in effect between the active and comparator treatments, and the expected standard deviation of the effects. These figures can be obtained from the literature reporting similar studies, or from pilot trials large enough to obtain sufficient data to estimate an effect size.

Because the trials are in effect small cluster randomized trials with the unit of randomization being the individual, adjustment of the sample size has to take into account the highest possible intra-cluster coefficient of one. The sample size has to be inflated accordingly. Computer modeling can be used to estimate sample size by running multiple simulations of the theoretical trial.

Analyzing the Data

This is dealt with in detail in Chap. 16.

Standard frequentist statistical techniques can be used to analyze the data. Because of the small numbers in the trials, there are numerous potential problems in applying these techniques. Most relate to the possibility of erroneous conclusions arising from the small sample size. A curious potential problem is the risk that confidence intervals may fall outside normality. For example, the lower limit of the confidence interval for changes in blood pressure might be a negative number.

The use of Bayesian statistics can avoid most of these problems. Bayesian statistics express the results as the likelihood that the observed result is true, and have a range from 0 to 1. The analysis uses prior evidence of an effect size, available through previous published work or pilot data to determine a clinically important difference between groups, and with successive trial data added, generates a posterior probability that changes with the addition of more trial data. A final estimate of effect is thus created when all the data are considered. As well as generating the effect size for the population under study, this estimate can be used to examine an individual's findings and make a judgment about whether the treatment was effective or not. Given that probabilities cannot be negative, it is not possible to have a credible interval less than 0 or greater than 1. The results are always credible. Techniques to analyze normally distributed and non-normally distributed data have been described (Nikles et al. 2011).

Conclusion

It may be very difficult to generate RCT evidence in populations with a low prevalence condition, or where it is difficult to recruit or retain participants in a study. Normal RCTs are very resource intensive in the latter groups, and practically impossible in the former setting. Aggregated N-of-1 trials open fresh opportunities to generate high quality evidence about treatment effects in populations like these, thereby increasing the likelihood that evidence based treatments will guide clinical practice in these settings.

References

Bland M (1995) An introduction to medical statistics. Oxford University Press, New York

Davis MP, Mitchell GK (2012) Topics in research: structuring studies in palliative care. Curr Opin Support Palliat Care 6:483–489

Hardy J, Quinn S, Fazekas B, Plummer J, Eckermann S, Agar M, Spruyt O, Rowett D, Currow DC (2012) Randomized, double-blind, placebo-controlled study to assess the efficacy and toxicity of subcutaneous ketamine in the management of cancer pain. J Clin Oncol 30:3611–3617

Mitchell G, Hardy J, Nikles J, Carmont S, Senior H, Schluter P, Good P, Currow D (2015) The effect of methylphenidate on fatigue in advanced cancer: an aggregated n-of-1 trial. J Pain Symptom Manag (in press). doi: 10.1016/jpainsymman.2015.03.009

Nikles J, Mitchell GK, Schluter P, Good P, Hardy J, Rowett D, Shelby-James T, Vohra S, Currow D (2011) Aggregating single patient (n-of-1) trials in populations where recruitment and retention was difficult: the case of palliative care. J Clin Epidemiol 64:471–480

Shelby-James TM, Hardy J, Agar M, Yates P, Mitchell G, Sanderson C, Luckett T, Abernethy AP, Currow DC (2012) Designing and conducting randomized controlled trials in palliative care: a summary of discussions from the 2010 clinical research forum of the Australian Palliative Care Clinical Studies Collaborative. Palliat Med 26:1042–1047

Chapter 6
Methodological Considerations for N-of-1 Trials

Keumhee C. Carriere, Yin Li, Geoffrey Mitchell, and Hugh Senior

Abstract N-of-1 trials are extremely useful in subject-focused investigations, for example, medical experiments. As far as we are aware, no guidelines are available in the literature on how to plan such a trial optimally. In this chapter, we discuss the considerations when choosing a particular N-of-1 trial design. We assume that the outcome of interest is measured on a continuous scale. Our discussion will be limited to comparisons of two treatments, without implying that the designs constructed can apply to non-continuous or binary outcomes. We construct optimal N-of-1 trials under various models depending upon how we accommodate the carryover effects and the error structures for the repeated measurements. Overall, we conclude that alternating between AB and BA pairs in subsequent cycles will result in practically optimal N-of-1 trials for a single patient, under all the models we considered without the need to guess at the correlation structure or conduct a pilot study. Alternating between AB and BA pairs in a single trial is nearly robust to misspecification of the error structure of the repeated measurements.

Keywords N-of-1 trials • Optimal design • Clinical trials • Crossover design • Residual effects • Direct treatment effect • Error structure • Adaptive trial design

Introduction

Medicine is an ever-changing science. Clinical trials are employed to conduct biomedical studies on human subjects to obtain specific answers about the impact of drug therapy and related medical interventions or treatments, generating efficacy data. The majority of such studies employ randomized controlled trials (RCTs)

K.C. Carriere (✉) • Y. Li
Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, Canada
e-mail: kccarrie@ualberta.ca; yin8@ualberta.ca

G. Mitchell
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: g.mitchell@uq.edu.au

H. Senior
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: h.senior@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_6

(Armitage 1975; Kenword and Jones 1987; Wei and Durham 1978). One approach to designing an RCT is the use of optimal experimental designs. An optimal design is a technique designed to assist a decision maker in identifying a preferable choice among many possible alternatives. Among the many RCT designs available, the most useful and popular design is the crossover design. For example, in a survey done in 1980 of numerous studies on the effects of antianxiety drugs on human performance, 68 % of the studies used the crossover approach (Brown 1980). It is still certainly one of the most popular approaches being adopted in many epidemiologic and pharmaceutical trials (Figueiras et al. 2005).

To illustrate the logistics of choosing a particular design, we first note that there are a number of excellent articles on optimal designs in the RCT literature. See for example Cheng and Wu 1980; Carriere 1994; Carriere and Huang 2000; Liang and Carriere 2009; Laska and Meisner 1985; (Afsarinejed and Hedayat 2002; Kunert and Stufken 2002, 2008). However, most of these designs, if not all, focus on optimizing the treatment effect for an average patient. The average patient is a construct – a virtual person who responds to the intervention by the mean of the population's responses. Individuals enrolled in a trial will respond better or worse than, or simply differently from the average patient. The available optimal designs are not adequate when studying individual-based treatment effects is desired.

Multi-crossover single-patient trials, known as N-of-1 trials, are often employed when the focus is to make the best possible treatment decision for an individual patient. From a clinician's perspective, having clear evidence of the value of one treatment over another (or no treatment) is far more useful than knowing the average response. The average response gives the clinician the *probability* that a treatment will be effective, whereas N-of-1 trials give far more certainty about whether the treatment for the patient sitting in front of them will work or not.

The simplest two-treatment N-of-1 trial uses the AB (or BA) sequence for treatments A and B; this treatment sequence has one crossover pair over two treatment periods. Each period is chosen to be of sufficient length for the treatments being tested to show an effect. Two periods (such as AB or BA) constitute a single cycle in a N-of-1 trial. As the patient becomes his or her own control, N-of-1 trials provide individual-based clinical evidence for the treatment effect, free of between-patient variations. With the rising cost of patient care, N-of-1 trials have the potential to be extremely useful, as they can minimize clinic visits and time on suboptimal treatments (Greenfield et al. 2007; Guyatt et al. 1986; Kravitz et al. 2004; Larson 1990; Nikles et al. 2005). To obtain stable estimates of the treatment effect, we desire to replicate such evidence for each patient. The question is then how many such cycles are desirable and what is the optimal order of treatment administration. Quite naturally, designing an N-of-1 trial involves deciding on the number of cycles and proper sequencing of treatments in order to plan the study optimally to achieve the trial objectives. The literature is lacking in providing these guidelines for constructing optimal N-of-1 trials. A recent book on N-of-1 trials also leaves the choice of an ideal design to the clinician, while suggesting various possible designs to consider (Kravitz et al. 2014).

The Number of Cycles and Sequences

For two-treatment experiments, a general N-of-1 trial can have multiple AB or BA crossover pairs called a cycle, in a sequence of treatments for within-patient comparisons. Hence, the number of periods, *p*, is a multiple of 2 and it is desirable to have more than two of these pairs or cycles for a stable estimate of a treatment effect. Typically, possible sequences to consider rapidly increase with increasing *p* in multi-period designs. For example, N-of-1 trials with three cycles and therefore $p=6$ would require considering $2^6 = 64$ treatment sequences before we could determine the optimal treatment sequence (s). Since it is only feasible to use a small set of cycles, we aim to determine the optimal sequences, while ensuring that each pair of periods consists of two distinct treatments. Therefore, such sequences as AAABBA or AAAABA are unlikely to be used in N-of-1 trials as AA are not two distinct treatments. In addition, we may consider block effects due to AB or BA, treating these as independent entities, which makes those sequences unsuitable to consider. Eliminating unsuitable sequences among 2^p for N-of-1 trials, we are left with $2^{p/2}$ distinct sequences to consider for *p*-period 2-treatment N-of-1 trials for *p/2* cycles. For example, a study with 6 periods (3 cycles) will be left with 8 distinct sequences to consider for a 6-period 2-treatment N-of-1 trial. In this example, an individual patient would be randomized to only 1 of the 8 possible sequences of treatment order.

There are many optimal repeated measurement designs available in the literature (Cheng and Wu 1980; Hedayat and Yang 2003). However, due to the special conditions mentioned above, these N-of-1 trial designs cannot be optimally derived from the existing designs. It is known that the two-treatment design (*AB, AA*) and their duals (*BA, BB*) is found to be universally optimal for two-period experiments, with the duality defined as the sequence that switches *A* and *B* with the same effect. Similarly, it is known that the two-sequence design *ABB* and its dual *BAA* and the four-sequence design (*ABBA, AABB*) and their duals (*BAAB, BBAA*) are optimal for three- and four-period experiments, respectively (Carriere 1994; Laska and Meisner 1985).

Straight application of this two-treatment optimal design literature with *A* to AB and *B* to BA would suggest that optimal N-of-1 trials can use the 4 sequence design with ABBA, ABAB and their duals for two within-patient comparisons, the 2 sequence design with ABBABA and its dual for three within-patient comparisons, and the 4 sequence design with ABBABAAB, ABABBABA and their duals for four within-patient comparisons. It is not yet known whether all of these sequences are indeed optimally equivalent so that each sequence is optimal for each individual patient for 4, 6 and 8-period N-of-1 trials. Further, applying the results from the literature would require at least two patients to utilize these existing designs, as the optimal design uses at least two sequences and is unsuitable for N-of-1 trials. In this Chapter, we show that not all sequences in these repeated measurement designs are optimal for N-of-1 trials for estimating individual-based treatment effects.

Ideally, when aggregated, the series of N-of-1 trials that are optimal for individual patients can also provide an optimal estimate of the treatment effects for the average patient. For example, in a multi-clinic setting in three AB pair six-period N-of-1 studies, all eight possible sequences ($2^{6/2} = 8$) have been used, i.e., ABABAB, ABABBA, ABBAAB, ABBABA and their duals to estimate both individual- based and average treatment effects (Guyatt et al. 1990). However, it is not known whether each of these eight sequences is optimal for individual patients. Further, it is not known whether a collection of the optimal and not-so-optimal N-of-1 trials will lead to optimal designs for estimating the average treatment effects. In the next Sections, we discuss how these do not lead to optimal aggregated N-of-1 trials for estimating the treatment effects for the average patient. We first discuss issues arising due to the repeated nature of these experiments.

Residual Treatment Effect

The main attraction of crossover designs is that the subject provides their own control, as measurements are taken repeatedly from the same subject using different treatments. If the treatment effects lasting beyond the given period are equal, they can provide efficient within-subject estimators of direct short-term treatment effects by removing between-subject variations.

However, there is one critical issue plaguing these repeated measurement designs and preventing them being popular despite their practical appeal. It is because they suffer from a long-standing controversy regarding residual treatment effects that last beyond the given period. N-of-1 trials are no exception. Sometimes referred to as the carryover effect, the residual effect is the effect of a previous treatment that carries over into the subsequent treatment periods. Thus, the effect of a treatment in a given period can be carried over to influence the responses in a subsequent period. Often, the residual effect of a treatment may be ignorable after two periods. However, the residual effect between the responses over two consecutive treatment periods may not be assumed to be negligible. A "washout" period placed between treatment periods could reduce the carryover effects, but a long washout period may increase the risk of drop-outs. Also, there is no guarantee that it completely removes the residual effects. Therefore, careful planning is important, as the nature of the carryover effect may be such that the N-of-1 trial method is not feasible (Bose and Mukherjee 2003; Carriere and Reinsel 1992; Kunert and Stufken 2002).

Nevertheless, the presence of residual effects does not invalidate the use of crossover designs. Rather it is the inequality of the residual effects of each treatment that may be causing the controversies. If the residual effects are equal for the treatments, then this has an effect in a statistical sense as if the residual effects do not exist, because they cancel out mathematically.

Despite the concern over residual effects, ethicists apparently have less of a problem with self-controlled designs such as crossover designs than with completely randomized or parallel-group designs (Carriere 1994). For example, when a trial involves treatments for patients with life-threatening conditions, it is almost beneficial to adopt these self-controlled designs. In parallel-group placebo-controlled studies, a group of sick patients is randomly allocated to a placebo or active group, with a 50 % chance of receiving a placebo or ineffective treatment for one or more periods of time, making them unethical for obvious reasons.

The way this is overcome in researching critical conditions depends on which of the following scenarios the researcher is addressing; either

1. The test treatment is given in addition to a proven or acceptable treatment in order to avoid a patient having to receive placebo alone, or
2. If the best known treatment has been shown to fail, only then considering a placebo-controlled design.

In the first scenario, the trial is: best treatment plus test intervention, compared with best treatment plus placebo. The second scenario is: test treatment vs placebo in a constrained population. Ethically either is acceptable. However, both alternatives raise interesting practical questions. The first is that if the proven treatment is effective, then the test treatment can only offer incremental effects. These are likely to be smaller, and therefore a larger sample size will be required to detect a clinically significant difference. Further, this design will not give information about the efficacy of the treatment on its own in the condition under consideration. In the alternative case of testing in the presence of failed proven treatments for a subset of patients, the potential participant pool will be far smaller, which may cause the proposed trial to be impractical to perform. Of course, test treatment vs placebo trials are ethically acceptable in critical conditions where there is currently no proven effective treatment.

In the literature, various models have been proposed to accommodate carryover effects. We will introduce and discuss the two most popular models in the next Section: (A) a model with a first-order residual effect and (B) a model with self and mixed carryover effects.

Models for N-of-1 Trials to Account for Carryover Effects

The most widely read statistical paper on the use of crossover experiments in clinical trials was published in 1965 by Grizzle, where the responses are modeled as:

Response = overall mean
+ period effects
+ sequence effects
+ direct treatment effects
+ residual treatment effects
+ measurement error.

Aside from the obvious overall mean effect and period effects, sequence effects may be present due to treatments given in a different order to patients, because some patients will be given AB or BA or some other order. While the primary objective is to study the direct treatment effects, their effects may not be unique due to the unequal residual treatment effects. Crossover design models have typically assumed that the treatments assigned to subjects have lasting effects on their responses to treatments in subsequent periods. A two-step approach has been used quite extensively where the unequal residual effects are first estimated and tested for significance before proceeding to estimate the direct treatment effects (Carriere 1994; Kunert and Stufken 2008).

Model with a First-Order Residual Effect

When it is assumed that the carryover effects last for only one period, this is known as a first-order residual effect model. In such a model, no interaction is assumed between the treatment administered during the current period and the carryover effects from the previous period. This interaction gives rise to the second-order residual effect. Hence, the model under this assumption is basically equal to the Eq. (1), where the term for the residual treatment effects is just the first-order residual effects, which last for only one more period of the treatment administration.

Model with Self and Mixed Carryover Effect

Taking the treatment and period interactions into account, Kunert (Kunert and Stufken 2002, 2008) presented an alternate model with self and mixed carryover effects. The self carry-over effect occurs when the treatment administered in the current and the previous period is the same; alternatively, if two different treatments are administered between the periods, it is known as a mixed carryover effect. The model under this assumption is more elaborate. From the Eq. (1), the term for the residual treatment effects will be split and be replaced with the following two terms:

- +Self carryover effects (if a preceding treatment is the same as the current one)
- +Mixed carryover effects (if the preceding treatment is not the same as the current one).

Optimal designs are highly model dependent. These effects are assumed to exist unless proven insignificant and therefore a reasonable effort should be made to separate them for an unbiased estimation of the direct treatment effects. Sometimes, however, it is simply practically impossible to accommodate all effects in the model. With N-of-1 trials, not all of these effects can be included in the model. Because we are dealing with just one subject and p responses in total, the period effects cannot be accommodated.

Repeated responses from a subject can be correlated and also involve measurement errors. The most popular structures may be to consider the pair of measurements as being equally correlated. Such a simple structure may work in designs with small numbers of periods. In N-of-1 trials with at least 4 periods, we may need to consider an auto-regressive structure, as the correlations may diminish gradually as the pair of measurements comes farther apart in their treatment periods. Here, the correlation is assumed large for a pair of measurements from two adjacent periods and decreases as the time between the two period increases. Some modification is possible by assuming them to be uncorrelated if they are more than two periods apart. See related discussion in Carriere (1994).

Design Parameters for N-of-1 Trials

In this section, we characterize N-of-1 trials in a small number of design parameters. Within each crossover pair AB and BA, the two treatments are to be different but, for two consecutive crossover pairs, the treatments assigned to the second period in the previous pair and the first period in the latter pair can be different or the same.

If an AB pair is followed by a BA pair, as in ABBA (or its dual, BAAB), we refer to the design as having an alternating pair in the sequence. Therefore, the sequence ABAB does not have an alternate cycle. The performance of an N-of-1 trial is related to how the pairs AB and BA alternate. By denoting s as the number of AA and BB and m as the number of AB and BA in a treatment sequence, we define $h = s - m$. There are $p - 1$ subsequences with a length of 2 in a p-period sequence. For example, for N-of-1 design with $p = 8$, ABABBABA, there are seven subsequences AB, BA, AB, BB, BA, AB, and BA. Here, $s = 1$ and $m = 6$, and $h = -5$. Table 6.1 shows the relationship between these design parameters for each of the eight possible eight-period sequences.

In the next section, we will see that sequences with the same h values have the same design properties. For instance, all three sequences ABABBAAB, ABBAABAB, ABBABAAB and their duals have $h = -3$, and contain the same amount of information about parameters of interest for this design. Hence, if this is the optimum value, the 8-period N-of-1 trial can use any one of these three sequences and their duals. Here, we immediately see from Table 6.1 that not all of the 8 sequence N-of-1 trials used elsewhere are optimal N-of-1 trials. Therefore, it is important to deliberate what the optimal N-of-1 trials are by examining how these cycles and pairs are organized.

Table 6.1 Sequences for $p = 8$ with corresponding design parameter values

h	Sequence	Alternation	s	m
−7	ABABABAB	0	0	7
−5	ABABABBA			
−5	ABABBABA	1	1	6
−5	ABBABABA			
−3	ABABBAAB			
−3	ABBAABAB	2	2	5
−3	ABBABAAB			
−1	ABBAABBA	3	3	4

Note: s is the number of AA and BB and m is the number of AB and BA in a treatment sequence with $h = s\text{-}m$

Optimal N-of-1 Trials

Typically, we are interested in estimating the direct and carryover treatment effects, while all others are treated as nuisance and secondary parameters. We build designs to this end. Due to the special nature of the design and the correlated data, we first consider how one could approach the data analysis. When data are obtained from an N-of-1 design, there are various ways to approach data analyses in order to gather any beneficial treatment evidence for the patient.

One is to observe effects between successive trial conditions, as shown in Fig. 6.1. Once the series of differences are observed, it would involve the usual repeated measures data analytic technique to plan and analyze them (Cantoni 2004; Davidian et al. 2009; Liang and Zeger 1986).

Alternately, one can observe and analyze treatment differences from each pair, as shown in Fig. 6.2. The same strategy of a longitudinal and repeated measures data analysis method as the first approach applies here, as the N-of-1 trial data were replaced by differences from successive pairs. The paired differences are simply regarded as repeated measurements, and more data simply means a better precision for the treatment effect (Carriere and Huang 2000). Such repeated measurements can be analyzed using various parametric and nonparametric methods (Liang and Zeger 1986; Senn 2002).

The third way to look at these N-of-1 trials data is similar to the first approach, but differs in that it holds the judgment or decision of a beneficial treatment effect for the given patient till the end of the trial, by using likelihood methods (Kravitz et al. 2014). As it analyzes the entire data set based on the employed model, the Type I error probability is minimized while possibly making multiple interim analyses and decisions. This is another consideration, discussed in this chapter in the development of optimal N-of-1 trial design. Such a model-based general approach could be more efficient than the first two approaches analyzing each cycle separately within each patient (Carriere 1994; Kravitz et al. 2014).

In this chapter, we recognize that N-of-1 designs deal with small samples, and thus we discuss an optimal strategy to collect the data based on the model by constructing an optimal design rather than specific data analysis strategies. To do so, we first need to define the information matrix under a particular model, which will contain information about all parameters of interest under the assumed model. More details

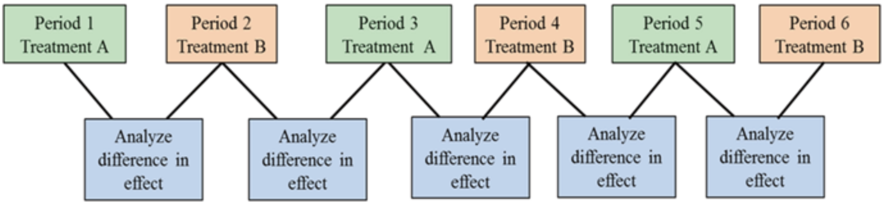

Period 1
Treatment A
Period 2
Treatment B
Period 3
Treatment A
Period 4
Treatment B
Period 5
Treatment A
Period 6
Treatment B
Analyze difference in effect
Analyze difference in effect
Analyze difference in effect
Analyze difference in effect
Analyze difference in effect

Fig. 6.1 Analysis of successive trial conditions in a six period N-of-1-design

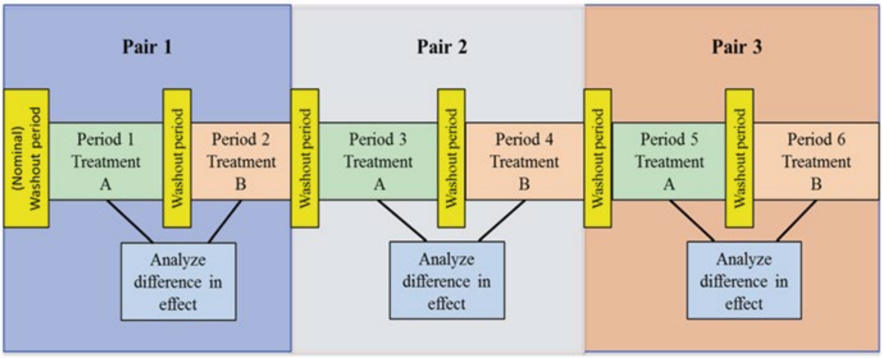


Fig. 6.2 Analysis of results from each successive pair in a six period of N-of-1 design

about an information matrix are found in much of the design literature, for example, Carriere and Reinsel (1993). Then, the optimal design for a set of parameters of interest is constructed using various optimality criterion. We use the D-optimality, which finds the optimal design by maximizing the determinant of the information matrix. See also Cheng and Wu (1980) and Kiefer (1975). Carriere (1994) and Carriere and Huang (2000) describe a practical approach to find optimal designs.

For N-of-1 trials, the optimal sequence is completely determined by h, as noted in the previous section and therefore, is much simpler to construct than previously. One could also find the optimal design that simultaneously optimizes all parameters of interest or the one that optimizes the carryover effects under the constraint that it optimizes the direct treatment effects. We note that the approach can also apply to find optimal designs for estimating some linear combinations of the parameters of interest. See also Carriere and Reinsel (1993). Since we are primarily interested in the optimal estimation of the direct treatment effects, we do not consider these cases. We summarize important results based on a D-optimality criterion in the next subsections (Carriere and Reinsel 1992).

Under the Traditional Model

Under the traditional model, we consider the error structure to be equally correlated between two measurements within the patient. We find the optimal design to consist of pairs of AB and BA appearing alternatively throughout the trial. Therefore, we find the following result.

Result 1 The optimal N-of-1 trial for estimating both the direct and residual effects is the one sequence design that consists of pairs of AB and BA appearing alternately.

For example, the optimal designs for N-of-1 trials with 4, 6, and 8 periods are the one sequence designs, ABBA, ABBAAB, and ABBAABBA, respectively. One could switch A and B to obtain a dual sequence with the same effect.

Repeated measures may be correlated highly when they are close together in time, while they may be negligible when they occur at a distance in time. One possible model for such a situation is to consider auto-regressive errors, accounting for correlation of measurements between two adjacent periods to be stronger than those far apart. It turns out that to discuss designs with auto-regressive errors, we need to consider whether the treatment given in the first and last periods are the same or not. However, the optimal designs are still determined by the value of *h*, and in general, the optimal design is to alternate AB and BA cycles as above.

Under the Model with Self and Mixed Carryover Effects

We also considered the model with self and mixed carryover effects. The optimal design was constructed by obtaining information matrices for the relevant parameters. Unlike the previous case, the optimal designs for estimating the direct treatment effect are not the same as those for residual effects.

The optimal design for estimating the direct treatment effect is the sequence with only AB pairs, such as ABABABAB. Although it may be of less interest, the optimal sequence for estimating the self carryover effect is to alternate between AB and BA pairs, while the optimal sequence for estimating the mixed carryover effect is to repeat the AB pair with no alternation. We summarize our findings in Result 2.

Result 2 The optimal N-of-1 trial for estimating the direct treatment and mixed carryover effect is the sequence with only AB pairs with no alternation, such as ABABABAB, while the optimal N-of-1 trial for estimating the self carryover effect is the sequence with AB and BA alternating throughout the trial.

Numerical Comparison

Although we constructed the optimal designs, it would be of interest to determine the practical benefit of having followed specific guidelines in adopting the optimal clinical trial design. To appreciate the practical performance of the optimal N-of-1 trials we constructed, we compare the efficiencies of some selected designs in estimating the treatment and carryover effects under the two models. We limit the comparison to the cases with independent and equi-correlated errors.

Recall that the optimal N-of-1 trials are either to alternate between AB and BA pairs or simply to repeat AB pair in a sequence. Under the traditional model, the optimal N-of-1 trial uses ABBAAB and ABBAABBA for 6 and 8 period experiments, respectively. We refer to them as S63 and S83. Under the self and mixed effects model, the optimal N-of-1 trial is to use ABABAB and ABABABAB for 6 and 8 period experiments, respectively, which we refer to as S61 and S81. Some other sequences are also considered, as defined in Table 6.2.

Table 6.2 Numerical illustration of various 6- and 8-period N-of-1 trials

		Traditional		Self/mixed		
Sequence	*h*	var(t)	var(g)	var(t)	var(s)	var(m)
S61: ABABAB	−5	1.208	1.500	1.208	NE	1.500
S62: ABABBA	−3	0.242	0.300	1.250	3.000	1.500
S63: ABBAAB	−1	0.173	0.214	1.214	1.714	1.714
S64: ABBABA	−3	0.242	0.300	1.250	3.000	1.500
S81: ABABABAB	−7	1.146	1.333	1.146	NE	1.333
S82: ABABBABA	−5	0.229	0.267	1.167	2.667	1.333
S83: ABBAABBA	−1	0.127	0.148	1.150	1.600	1.400
S84: ABBABAAB	−3	0.150	0.174	1.147	1.647	1.412

Note: Entries are relative variances for a direct treatment effect *t*, a first-order carryover effect *g*, a self carryover effect *s*, and a mixed carryover effect *m,* under the traditional and self/mixed carryover effects models. S61-S64 are 6-period N-of-1 trials and S81-S84 are 8-period N-of-1 trials with the order of cycles as given

Table 6.2 tabulates the variances of the treatment effects for 8 period N-of-1 trials under the two models. Table 6.2 reveals that the effects of choosing a specific sequence under the self and mixed carryover effect model are rather minimal. Careful examination also reveals that the optimal sequence for all three (direct and residual carryover) treatment effects simultaneously is the same as that under the traditional model.

Specifically, Table 6.2 shows that the optimal individual-based N-of-1 trials are S63 and S81 for estimating the direct treatment effects under respective models, as expected. However, there are no real practical differences among various N-of-1 trials under the self and mixed model. Under the self and mixed carryover model, although S81 repeating AB pairs in each cycle is optimal for estimating the direct treatment effect, it does not allow estimation of self carryover effects, making S63 and S83 preferable. Therefore, for robust and optimal N-of-1 trials it is recommended to use a sequence alternating between AB and BA pairs, such as S63 and S83, under all models.

In summary, it appears that there is no discernable or sizable advantage to distinguish among the two models and possibly various error structures. The above comparison remains true under both independent and equi-correlated error structures.

Overall, S63 and S83 for single N-of-1 trials seem to be the best under both models. They are optimal for estimating direct treatment and mixed carryover effects. Further, they are optimal for estimating both the treatment and carryover effects under the traditional model.

Table 6.2 also shows that increasing the number of periods from 6 to 8 will result in an efficiency gain of 0.173/0.127 = 36 % under the traditional model, while the gain is not as substantial under a complex model.

In general, we suggest that alternating AB and BA pairs in sequence will result in a nearly optimal design, if not the optimal one, under all models we considered, for estimating individual effects in N-of-1 trials.

Adaptive Trial Design

Adaptive designs are gaining popularity in recent years. Liang and Carriere (2009) outline how one could plan response adaptive designs utilizing outcomes in a given experiment while achieving multiple objectives. For example, clinicians may wish to achieve good estimation precision, effective treatment of patients, or cost effectiveness (Carriere and Huang 2000). Recently, Liang, Li, Wang, and Carriere (Liang and Carriere 2009) extended their approach to binary responses. For N-of-1 trials, designs can be found by updating AB or BA pairs successively as the trial progresses. Such objectives as maintaining a balance or counterbalancing between AB and BA pairs as suggested by Kravitz et al. (2014) can also be considered. A Bayesian framework may be the most natural for adaptive design and decision-making. However, not much attention has been given to finding optimal designs for binary data in the literature, and further research is needed.

Discussion

N-of-1 trials are extremely useful in subject-focused investigations, for example, medical experiments. As far as we are aware, no guidelines are available in the literature on how to plan such a trial optimally. In this Chapter, we construct optimal N-of-1 trials under various models depending upon how we accommodate the carryover effects and the error structures for the repeated measurements. We also suggest constructing optimal aggregated N-of-1 trials for both the individual and average patients.

A straight application of the two-treatment optimal design results in the literature with *A* to AB and *B* to BA can result in an inferior design unsuitable for individual patient treatment care. We showed that not all of the suggested sequences are optimal to be used in N-of-1 trials. Further, they may not be optimal for estimating effects at the average level. For example, when $p=4$, the literature gives *ABBA* and *AABB* and their duals as the optimal design. Applying this result to 8-period N-of-1 trials, we would need to consider at least four sequences, ABBABAAB, ABABBABA and their duals. However, we showed that none of these four sequences are optimal for N-of-1 trials for $p=8$.

For the traditional first-order residual effects model with uncorrelated and equal-correlated errors, the optimal N-of-1 trial design is to use the sequence consisting of alternating AB and BA pairs. We can use its dual sequence with the same effect. For example, the optimal 8 period N-of-1 trial design is to use ABBAABBA or BAABBAAB under equal or uncorrelated errors. For the self and mixed effect model, the optimal N-of-1 trial uses the sequence consisting of only AB pairs. If self and mixed carryover effects are a concern, the optimal 8 period N-of-1 trial is ABABABAB for estimating the direct treatment effects. Hence, once again we find that optimal designs are strongly model dependent. It would depend on how we accommodate the residual effects and what type of errors are practically feasible.

Surprisingly, however, the results do not change by very much in practice with adopting not so optimal sequences, as we examined in Section "Optimal N-of-1 Trials". A numerical calculation of the estimation precision using several 6- and 8-period designs revealed the actual performance of a particular design, giving us practical guidelines. Overall, we conclude that alternating between AB and BA pairs in subsequent cycles will result in practically optimal N-of-1 trials for a single patient, if not the optimal, under all the models we considered without the need to guess at the correlation structure or conduct a pilot study. Alternating between AB and BA pairs in a single trial is nearly robust to misspecification of the error structure of the repeated measurements.

Lastly, we suggest that when an experiment has been carried out with the optimal N-of-1 trial and additional patients are accrued in the trial, we can plan and aggregate these N-of-1 trials optimally by allocating the same number of patients to its dual sequence by reversing the treatment order, thereby optimizing the trial for both the individual and average patients.

Acknowledgments This research was made possible due to the funding from the Natural Sciences and Engineering Research Council of Canada.

References

Afsarinejed K, Hedayat A (2002) Repeated measurement designs for a model with self and mixed carryover effects. J Stat Plan Inference 106:449–459

Armitage P (1975) Sequential medical trials. Blackwell, Oxford

Bose M, Mukherjee B (2003) Optimal crossover designs under a general model. Stat Probab Lett 62:413–418

Brown BWJ (1980) The crossover experiment for clinical trials. Biometrics 36:69–79

Cantoni E (2004) A robust approach to longitudinal data analysis. Can J Stat 32:169–180

Carriere KC (1994) Cross-over designs for clinical trials. Stat Med 13:1063–1069

Carriere KC, Huang R (2000) Crossover designs for two-treatment cancer clinical trials. J Stat Plan Inference 87:125–134

Carriere KC, Reinsel GC (1992) Investigation of dual-balanced crossover designs for two treatments. Biometrics 48:1157–1164

Carriere KC, Reinsel GC (1993) Optimal two-period repeated measurements designs with two or more treatments. Biometrika 80:924–929

Cheng CS, Wu CF (1980) Balanced repeated measurements designs. Ann Stat 8:1272–1283

Davidian M, Verbeke G, Molenberghs G (2009) Longitudinal data analysis. Int Stat Rev 77:1857–1936

Figueiras A, Carracedo-Martinez E, Saez M, Taracido M (2005) Analysis of case-crossover designs using longitudinal approaches: a simulation study. Epidemiology 16:239–246

Greenfield S, Kravits R, Duan N, Kaplan SH (2007) Heterogeneity of treatment effects: implications for guidelines, payment, and quality assessment. Am J Med 120:S3–S9

Grizzle JE (1965) The two-period change over design and its use in clinical trials. Biometrics 21:461–480

Guyatt G, Sackett D, Taylor DW, Chong J, Roberts R, Pugsley S (1986) Determining optimal therapy–randomized trials in individual patients. N Engl J Med 314:889–892

Guyatt GH, Heyting A, Jaeschke R, Keller J, Adachi JD, Roberts RS (1990) N of 1 randomized trials for investigating new drugs. Control Clin Trials 11:88–100

Hedayat AS, Yang M (2003) Universal optimality of balanced uniform crossover designs. Ann Stat 31:978–983

Kenword M, Jones B (1987) A log-linear model for binary crossover data. Appl Stat 36:192–204

Kiefer J (1975) Construction and optimality of generalized Youden designs. In: Srivastava JN (ed) A survey of statistical designs and linear models. North-Halland, Amsterdam

Kravitz RL, Duan N, Breslow J (2004) Evidence-based Medicine, heterogeneity of treatment effects, and the trouble with averages. Milbank Q 82:661–687

Kravitz RL, Duan N (eds) and The Decide Methods Center N-of-1 Guidance Panel (Duan N, EI, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S) (2014) Design and implementation of N-of-1 trials: a user's guide. Agency for Healthcare Research and Quality, Rockville

Kunert J, Stufken J (2002) Optimal crossover designs in a model with self and mixed carryover effects. J Am Stat Assoc 97:896–906

Kunert J, Stufken J (2008) Optimal crossover designs for two treatments in the presence of mixed and self-carryover effects. J Am Stat Assoc 103:1641–1647 (correction 1999, 86, 234)

Larson EB (1990) N-of-1 clinical trials. West J Med 152:52–56

Laska E, Meisner M (1985) A variational approach to optimal two-treatment crossover designs: application to carryover-effect models. J Am Stat Assoc 80:704–710

Liang YY, Carriere KC (2009) Multiple-objective response-adaptive repeated measurement designs for clinical trials. J Stat Plan Inference 139:1134–1145

Liang KY, Zeger S (1986) Longitudinal data analysis using generalized linear models. Biometrika 73:13–22

Nikles CJ, Clavarino AM, Del Mar CB (2005) Using N-of-1 trials as a clinical tool to improve prescribing. Br J Gen Pract 55:175–180

Senn S (2002) Crossover trials in clinical research. Wiley, Hoboken

Wei LJ, Durham S (1978) The randomized play the winner rule in medical trials. J Am Stat Assoc 73:840–843

Chapter 7
Randomization, Allocation Concealment, and Blinding

Hugh Senior

Abstract Of major concern in N-of-1 trials, common to any epidemiological approach, is the introduction of bias and confounding. These factors may modify the size of the treatment estimate or shift the treatment estimate away from its true value. The methodological approaches of randomization, allocation concealment, and blinding are employed to prevent or minimize confounding and bias in trials. This chapter provides definitions and describes the various methods of randomization, allocation concealment, and blinding that can be adopted in N-of-1 trials. In addition, the chapter details the roles of specific research staff and the information required for the reporting of N-of-1 trial blinding methods in medical journals.

Keywords N-of-1 trial • Randomization • Allocation concealment • Blinding • Single-blind • Double-blind • Triple-blind • Bias • Confounding • Allocation sequence

Introduction

N-of-1 trials are cross-over trials with multiple cycles where each patient receives both the intervention and control treatments. Randomization in N-of-1 trials, as in other trial designs, aims to minimize confounding and selection bias. The process of randomization in N-of-1 trials involves the random selection of the order of treatments for each patient. Allocation concealment is a process of concealing the allocation sequence from the investigator responsible for recruiting patients. It prevents investigators from influencing the assignment of treatment to a patient and thus prevents selection bias. Blinding is the process of keeping investigators, patients and other researchers unaware of the assigned treatments within an N-of-1 trial to minimize ascertainment bias. Ascertainment bias occurs if the results and conclusions of a trial are influenced by knowledge of the trial medication each participant

H. Senior (✉)
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: h.senior@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_7

is receiving. Assessments of subjective outcomes are especially prone to ascertainment bias (Schulz and Grimes 2002b).

This chapter defines and describes the methods of randomization, allocation concealment and blinding in N-of-1 trials, along with the roles of key research staff and the minimum information required for reporting trial methods in relation to randomization, allocation concealment and blinding in medical journals.

Randomization in N-of-1 Trials

In N-of-1 trials, randomization aims to provide a sequence of treatments that are balanced to minimize confounders, including those that are time dependent, which could potentially modify the treatment estimate. Randomization not only controls for confounding, it also prevents selection bias by randomly selecting the sequence of treatments independent of the investigator and patient. A bias is a systematic error which can lead to an estimated treatment effect deviating in one direction from the true value (Friedman et al. 1998).

The typical N-of-1 trial design with two treatments is composed of three cycles, where in each cycle the patient is assigned a pair of treatments (e.g., AB or BA). Within each cycle, the order of treatments (e.g. drug A (intervention) or placebo B (control)) is randomly allocated and each cycle has identical time periods. This methodological approach ensures a counterbalance between treatments within each cycle and across the trial to minimize any imbalance in potential confounding factor(s) including those that are time-dependent. This is especially true of aggregated N-of-1 trials as time dependent confounding factors are more completely balanced across treatments, as compared to individual N-of-1 trials, due to all possible allocation sequences being incorporated into the model.

For a three cycle cross-over design which is counterbalanced within each cycle, where the order may be, for example, AB,BA,AB; eight allocation sequences are possible. Unlike traditional parallel randomized clinical trial where a patient is randomized to one of two treatment groups (A or B), in N-of-1 trials, the patient is randomized to an order of treatments, namely, the allocation sequence. Operationally, the study statistician would randomly select from the full set of allocation sequences using simple randomization methods to produce a randomization list for the trial. The statistician may employ tables of random numbers or random functions on statistical software or calculators to produce a randomization list (Roberts and Torgerson 1998).

The statistician would provide the randomization list to a colleague. Dependent on the allocation concealment method (see section below on allocation concealment), the colleague will either be a pharmacist, or data manager (e.g. if using centralized computer randomization) or an independent investigator (e.g. if allocation is provided using opaque envelopes).

When reporting the trial findings in published works, information on the method used to generate the random allocation sequence must be detailed in the manuscript (Schulz et al. 2010).

Allocation Concealment in N-of-1 Trials

Definition and Purpose of Allocation Concealment

Allocation concealment is “a technique used to prevent selection bias by concealing the allocation sequence from those assigning participants to intervention groups, until the moment of assignment. This procedure prevents researchers from (unconsciously or otherwise) influencing which participants are assigned to a given intervention group” (CONSORT 2015).

A selection bias is a bias introduced to the study outcomes if an investigator who recruits patients is made aware of the allocation sequence the patient will receive, and either uses this knowledge to exclude or discourage some patients from participating in the trial based on their prognosis or delays the patient’s allocation until a more preferable allocation sequence arrives (Elkins 2013).

Allocation Concealment Methods

Inadequate or unclear allocation concealment methods can lead to an inflated estimate of treatment effects of up to 40 %, potentially leading to biased findings and healthcare recommendations (Gluud 2006; Herbison et al. 2011; Juni et al. 2001; Kjaergard et al. 2001; Moher et al. 1998). In these traditional parallel group trial designs, inadequate concealment has led to selection bias, the bias that randomization aimed to prevent, as the allocation is no longer randomly assigned, leading to prognostic differences between groups. Consequently, the allocation sequence in an N-of-1 trial must be strictly concealed from the recruiting investigator to prevent any foreknowledge of the sequence. The risk of selection bias is less in an N-of-trial design compared to a traditional trial design as it uses a “within subject” study design, where each patient receives both the intervention and control. However, it is good research practice to ensure allocation concealment for all trial designs, and it is a requirement of the CONSORT statement on the reporting of trials (Schulz et al. 2010).

There are number of methods to provide adequate allocation concealment. These include the use of sequentially numbered envelopes, pharmacy controlled allocation, and central randomization (Schulz and Grimes 2002a).

The Envelopes method to conceal allocation employs “sequentially numbered opaque sealed envelopes” (SNOSE) which contain the allocation. The envelopes are prepared by an independent agent, sealed, and provided to the recruiting investigator(s) (Elkins 2013). However, the envelope method is criticized as being susceptible to manipulation by holding the envelopes to a bright light to see the next allocation, or opening the envelope prior to assigning the patient (Viera and Bangdiwala 2007; Hewitt et al. 2005). Additional safeguards to improve the envelope method are to (i) ensure envelopes are identical in appearance and weight, (ii) number the envelopes in advance, (iii) place pressure sensitive or carbon paper inside the envelope to record the patient's name onto an assignment card thereby creating an audit trail, (iii) insert a material such as cardboard or tinfoil in the envelope that is impermeable to bright light and (iv) ensure the envelopes are opened sequentially only after the patient's details are written on the face of the envelope (Schulz and Grimes 2002a; Viera and Bangdiwala 2007).

More trial designs including N-of-1 trials are using pharmacy controlled allocation to ensure allocation concealment. Eligible patients are registered into the trial by the recruiting investigator. The investigator provides the patient with a prescription for the trial medications, which the patient takes to the pharmacy. The pharmacist prepares and dispenses the trial medications in numbered containers (for example, week 1, week 2…) according to a pre-prepared numbered randomization list that lists the next allocation sequence. The pharmacist writes the patient's details alongside the next allocation sequence on the randomization list. This method also ensures the pharmacy can provide unblinding in the case of an emergency (see later in this chapter). In some trial designs such as community N-of-1 trials, the prescription may be emailed/faxed to the pharmacy, who will prepare the trial medication according to the randomization list and courier the medication containers to the patient at home, workplace or community pharmacy. Investigators must ensure the pharmacy is provided with and follows standard operating procedures for randomization and allocation concealment.

Centralized computer systems that can be accessed through a remote computer via the internet or by telephone (through an "Interactive Voice Response System") or fax/e-mail are a popular method of randomization and allocation concealment. This method is especially useful for multisite trials. This method ensures allocation concealment by only assigning the next allocation sequence for an individual if the patient is eligible and enrolled in the study. In an N-of-1 trial, the central computer will provide a number of a sequence which is next in the randomization list. This number is provided to the pharmacy who dispenses medications according to the sequence of medications assigned denoted to that number. In a common N-of-1 trial design where two treatments are randomly allocated in three cycles, for example, AB BA AB, there are $2^3 = 8$ possible sequences (namely, ABBAAB, ABABAB, ABBABA, ABABBA, BABAAB, BAABAB, BABABA, BAABBA). The pharmacy will be provided with a number from 1 to 8 by the centralized computer system with each number denoting a specific sequence. An additional advantage of the centralized computer system of treatment allocation is that the system also monitors allocation concealment though time stamps and electronic logs (Viera and Bangdiwala 2007).

Reporting Allocation Concealment

Reporting of the mechanisms of allocation concealment in journal articles has historically been poor. In evaluations of the quality of reporting, 50–70 % of trial reports did not include sufficient information to determine if the allocation concealment methods were adequate (Clark et al. 2013; Hewitt et al. 2005). The CONSORT statement on reporting of trials states that investigators must provide information on the mechanism used to implement the random allocation sequence including the description of any steps taken to conceal the sequence prior to assigning the trial medications (Schulz et al. 2010).

Blinding of N-of-1 Trials

Definition and Purpose of Blinding

Blinding (also known as "Masking") is a process that attempts to ensure that those who are blinded are unaware of the treatment group (e.g. active drug or placebo) for the duration of the trial (Schulz and Grimes 2002b). Those who are blinded may include participants, investigators, assessors who collect the data, data safety monitoring boards, and statisticians (Schulz and Grimes 2002b; Viera and Bangdiwala 2007).

Blinding ensures that the participants, investigators and assessors are not influenced by their knowledge of the intervention, thereby minimizing ascertainment bias (Schulz et al. 1995, 2002; Schulz and Grimes 2002b; Forder et al. 2005).

In N-of-1 trials, blinding reduces the likelihood that participants will bias any physical or psychological responses to therapy (e.g. quality of life or pain levels) due to their preconceived perception of the value of treatment, and the reporting of side effects (Schulz and Grimes 2002b; Matthews 2000; Friedman et al. 1998). As participants in an N-of-1 trial receive both the control and intervention therapies during the trial, concerns that do occur in traditional trials that are minimized by blinding may be less relevant. These include participants seeking adjunct therapies or withdrawing from the trial as they are dissatisfied with being randomized to a placebo control.

For investigators, blinding reduces the influence of the knowledge of the intervention on their attitude and advice to participants, patient management, their likelihood of differentially adjusting the dose or administering co-interventions, or differentially guiding participants to withdraw (Schulz and Grimes 2002b; Viera and Bangdiwala 2007; Matthews 2000). Blinding reduces the likelihood that assessors (which may be the investigator and/or another health professional) will differentially assess an outcome based on their knowledge of the assigned treatment (Viera and Bangdiwala 2007).

Please note that the term "investigators" in describing blinding is a term broadly assigned to the trial team, which may include among others the trial designers, trial recruiters, assessors, and health care providers treating the participant (Schulz and Grimes 2002b).

Types of Blinding

Blinded trials can be of three types, namely single or double or triple blind. The reader should note that these terms do not have clear definitions and are often reported incorrectly in the literature.

Commonly, in a single-blinded trial, the participant (or sometimes the investigator) is unaware of the treatment assignment, but everyone else involved is. In some cases, this can refer to a situation where the participant and the investigator know the treatment assigned, whereas the assessor is blinded (Schulz and Grimes 2002b). A single-blinded trial may be the best approach if there is a clear rationale that the investigator must keep the participant blind to reduce bias, but their knowledge is critical for the participant's health and safety (Friedman et al. 1998). Alternatively, if the intervention is actually delivered by a clinician, and the N-of-1 trial is for the purpose of guiding treatment decisions, it may not be practical to blind the clinician. The disadvantage of the single-blind is that the investigator may consciously or subconsciously bias the study by biasing data collection, or differential prescription of concomitant therapy or differential patient management (Friedman et al. 1998).

In a double-blinded trial, the participants, investigators and assessors are blinded throughout the trial (Schulz and Grimes 2002b). As already stated, a double-blind approach eliminates or minimizes the risk of bias in the trial. In a double-blind and a triple blind (see below) trial it is important that there is a procedures for blinding when assigning the interventions, and that a separate body who can be unblinded, such as an "Independent Data Monitoring Committee" (see Chap. 10 on Adverse Events), are responsible for assessing the data for any adverse effects and benefit (Friedman et al. 1998).

A triple-blinded trial has the same characteristics as a double-blind trial with the addition that those who adjudicate the study outcomes are also blinded (Schulz and Grimes 2002b; Forder et al. 2005). This can be achieved by ensuring the data monitoring committee are not provided with the identity of the groups, instead, they are assigned a code such as group A and B. This approach assumes the data monitoring committee could be biased if the randomization status was known to them in their assessment of adverse effects or benefit. Some investigators feel this may impede the ability of the committee to perform their tasks of safeguarding participants by looking at individual cases (Friedman et al. 1998). This is a decision which needs to be made during the design of the trial in consultation with the data monitoring committee. If a triple-blinded trial is chosen, the data monitoring committee should have the option to break the blind if the direction of an observed trend in group A or B requires a further unblinded investigation (Friedman et al. 1998). Some trials blind the study statistician to reduce bias during analysis again by assigning the dummy codes of A and B to the trial groups (Matthews 2000). The groups only become unblinded at the end of the analysis for reporting purposes.

Utilizing a double or triple blind is highly recommended in N-of-1 drug trials to determine drug efficacy.

The Role of the Placebo in Blinding and the Placebo Effect

A placebo is a pharmacologically inactive agent used in the control group of the trial (Schulz et al. 2002). It is used in trials where the investigator is not assessing the effectiveness of a new active agent against an effective standard agent. Indeed, it is more ethically sound not to provide a placebo control if an effective standard agent exists to act as the control. However, even if a standard agent is used as a control, investigators may include placebos by using the double-dummy method for blinding (see the following section). The use of a placebo agent is critical for achieving trial blindness.

The use of a placebo not only maintains a blind, but it also excludes the "placebo effect" in the trial. The placebo effect occurs when an inactive agent is administered to a patient, but this has a beneficial effect on the attitude of participants, thereby producing a response (Schulz and Grimes 2002b; Matthews 2000). The placebo effect occurs both in the control group and the intervention group, therefore, the provision of a placebo balances the placebo effect across the trial groups (Schulz and Grimes 2002b).

In some trials, investigators may involve an active placebo, rather than an inactive placebo. An active placebo contains substances that will produce the symptoms or side effects that will occur in the active intervention agent, thereby ensuring the blind is not broken as these effects otherwise would identify the active investigational agent (Schulz and Grimes 2002b). Most placebo-controlled trials use an inactive placebo.

Drug Matching for Placebos

To ensure a blinded trial, the placebo control agent and the intervention agent must be similar.

This is especially important in cross-over trials including N-of-1 trials where participants receive both the control (e.g. placebo) and intervention agents. The trial agents must be similar in appearance (size, shape, color, sheen, and texture) (Friedman et al. 1998). It may be necessary to also ensure that the taste (and odor) is the same using masking agents. It is good practice to pre-test the similarity of the trial agents by asking a group of independent observers not involved in the trial to see if they can observe any differences (Friedman et al. 1998).

The most common method for drug matching in trials, and to reduce trial costs, is to over-encapsulate the trial agents. Over-encapsulation is the process of placing trial tablets or capsules in a hard gelatin capsule and backfilling with an inactive excipient to produce identical capsules. When an investigator decides to produce drug matches by over-encapsulation they must consider the size of the final capsule and whether this will be difficult to swallow for participants who have swallowing difficulties, such as older people, stroke patients, and young children.

The Double-Dummy Method for Blinding Non-placebo Trials

If a new trial agent is to be compared against an effective standard agent as control, and these two agents are dissimilar in characteristics, blinding can still be achieved using the double-dummy method (Schulz and Grimes 2002b). The investigators will need to ask the pharmaceutical companies who manufacture the drug to supply matching placebos, or have matching placebos prepared by a pharmacy. If a participant is randomized to the intervention agent, they would be prescribed the intervention active agent and a control inactive placebo. Vice versa, if a participant is randomized to the control agent, they would be prescribed the control active agent and an intervention inactive placebo. This approach will enact a double blind as participants and investigators are unable to distinguish which agent is an active agent or an inactive placebo. For example, in a set of N-of-1 trials to compare the non-steroidal anti-inflammatory drug celecoxib with paracetamol for osteoarthritis, to maintain a blind, patients who were randomized to active celecoxib within a trial period, were administered active celecoxib and an inactive placebo that was identical to active paracetamol to be taken simultaneously. Alternatively, those randomized to active paracetamol, were administered active paracetamol and an inactive placebo that was identical to active celecoxib to be taken simultaneously (Yelland et al. 2007).

Assessing If Blinding was Successful

It is possible to assess the success of blinding by asking participants, investigators and assessors in a N-of-1 trial during each period of a cycle whether they believe the control or intervention agent was administered. If these individuals are more accurate than chance in identifying an agent, then the trial may not have a successful blind (Schulz and Grimes 2002b). In a common N-of-1 trial design where two treatments are randomly allocated in three cycles, for example, AB BA AB, there are $2^3 = 8$ possible sequences. Therefore, the chance an individual will be able to guess the exact sequence is one in eight.

Unblinding Procedures

All blinded trials must have unblinding processes for individual participants, especially for a "Data Monitoring Committee" to access benefit and safety of the trial, and also for doctors to be able to identify what agents an individual is prescribed in a medical emergency (Matthews 2000). In many cases, the participant can be withdrawn from the trial medication without breaking the blind. Unblinding procedures may involve an individual or group (e.g. trial pharmacy) other than the trial investigator to ensure the investigator and participant can remain blinded. If a participant or investigator is unblinded, this must be noted as a protocol deviation/violation. For additional information on unblinding procedures, see Chap. 10.

Providing Individual Trial Reports and Maintaining the Blind

One major advantage of N-of-1 trial designs over other trial designs is that at the end of the trial for an individual participant, the individual's data can be analyzed and a report prepared on the effectiveness of the intervention compared to control, for a clinician to consult with the patient on whether to continue with trial medication under routine care (see Chap. 9). However, if the study is a set of N-of-1 trials, a problem arises, as even though an individual patient has finished the trial and a report has been generated, other participants are still to be recruited or are still being followed in the trial by the same investigator. Clinicians could possibly perceive a pattern of treatment effect in certain sorts of individuals, whether that is a true observation or not, thereby risking ascertainment bias. The challenge is how to produce a report for an individual during a live trial while maintaining the blind of the remaining participants and the investigators. A procedure adapted by our research group is to have the unblinded statistician provide the individual analyses and send the report data to an independent academic clinician, who will prepare the report and send it to the patient's doctor for consultation. If the statistician is also blinded, the statistician can conduct an individual data analysis using dummy codes and send the blinded individual findings to the academic clinician not involved in the trial to unblind and prepare the individual patient's report. Another strategy is to have one investigator responsible for the analyzing and reporting the trial results to the patient, and a different investigator recruiting and conducting assessments for the trial.

Reporting of Blinding in Publications

For many journals, the extent of blinding must be reported according to the CONSORT statement on reporting trial findings (Schulz et al. 2010). Information must be provided on who was blinded, the methods of blinding, on what characteristics the control and intervention agents were matched, where the randomization schedule was held, if individuals or the trial were unblinded at any stage, and how the success of the blind was assessed (Schulz and Grimes 2002b; Viera and Bangdiwala 2007).

Conclusion

Investigators should dedicate adequate time and resources to prepare a trial protocol, prior to study commencement, that details the procedures of randomization, allocation concealment and blinding. In N-of-1 trials, this should also include the preparation of the individual patient report for the patient's doctor for consultation while maintaining the blind for those who remain in the trial.

The implementation of these procedures during the trial will ensure the prevention or minimization of confounding and bias, and accordingly their influence on the estimate of the treatment effect, to allow reporting of accurate and credible findings.

References

Clark L, Schmidt U, Tharmanathan P, Adamson J, Hewitt C, Torgerson D (2013) Allocation concealment: a methodological review. J Eval Clin Pract 19:708–712

Consort (2015) Consort glossary. Available: http://www.consort-statement.org/resources/glossary

Elkins M (2013) Concealed allocation in randomised trials. J Physiother 59:134–136

Forder PM, Gebski VJ, Keech AC (2005) Allocation concealment and blinding: when ignorance is bliss. Med J Aust 182:87–89

Friedman LM, Demets DL, Furberg C (1998) Fundamentals of clinical trials. Springer, New York

Gluud LL (2006) Bias in clinical intervention research. Am J Epidemiol 163:493–501

Herbison P, Hay-Smith J, Gillespie WJ (2011) Different methods of allocation to groups in randomized trials are associated with different levels of bias. A meta-epidemiological study. J Clin Epidemiol 64:1070–1075

Hewitt C, Hahn S, Torgerson DJ, Watson J, Bland JM (2005) Adequacy and reporting of allocation concealment: review of recent trials published in four general medical journals. BMJ 330:1057–1058

Juni P, Altman DG, Egger M (2001) Systematic reviews in health care: assessing the quality of controlled clinical trials. BMJ 323:42–46

Kjaergard LL, Villumsen J, Gluud C (2001) Reported methodologic quality and discrepancies between large and small randomized trials in meta-analyses. Ann Intern Med 135:982–989

Matthews J (2000) An introduction to randomized controlled clinical trials. Arnold, London

Moher D, Pham B, Jones A, Cook DJ, Jadad AR, Moher M, Tugwell P, Klassen TP (1998) Does quality of reports of randomised trials affect estimates of intervention efficacy reported in meta-analyses? Lancet 352:609–613

Roberts C, Torgerson D (1998) Randomisation methods in controlled trials. BMJ 317:1301

Schulz KF, Grimes DA (2002a) Allocation concealment in randomised trials: defending against deciphering. Lancet 359:614–618

Schulz KF, Grimes DA (2002b) Blinding in randomised trials: hiding who got what. Lancet 359:696–700

Schulz KF, Chalmers I, Hayes RJ, Altman DG (1995) Empirical evidence of bias. Dimensions of methodological quality associated with estimates of treatment effects in controlled trials. JAMA 273:408–412

Schulz KF, Chalmers I, Altman DG (2002) The landscape and lexicon of blinding in randomized trials. Ann Intern Med 136:254–259

Schulz KF, Altman DG, Moher D, Group C (2010) CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332

Viera AJ, Bangdiwala SI (2007) Eliminating bias in randomized controlled trials: importance of allocation concealment and masking. Fam Med 39:132–137

Yelland MJ, Nikles CJ, Mcnairn N, Del Mar CB, Schluter PJ, Brown RM (2007) Celecoxib compared with sustained-release paracetamol for osteoarthritis: a series of n-of-1 trials. Rheumatology (Oxford) 46:135–140

Chapter 8
Data Collection and Quality Control

Hugh Senior

Abstract To achieve a reliable data set for analysis that complies with the protocol, a system of clinical data management (CDM) is critical. CDM is the planning and process of data collection, integration and validation. This chapter provides a synopsis of the key components of CDM which need to be considered during the design phase of any trial. Topics addressed include the roles and responsibilities of research staff, the design of case report forms for collecting data; the design and development of a clinical database management system, subject enrolment and data entry, data validation, medical coding, database close-out, data lock and archiving. An additional section discusses the rationale behind trial registration.

Keywords Data management • Validation • Case report forms • Trials • Missing data • Database • ICH-GCP • Enrolment • CONSORT • Adverse events • Source data • Data validation • Logic check • Data discrepancies • Medical coding • Trial registration

Introduction

Data management is an essential component in clinical trials to ensure that data that are analyzed are reliable and statistically sound (Krishnankutty et al. 2012). Clinical data management (CDM) aims to ensure high quality data by minimizing errors and missing data to ensure a reliable dataset for analysis (Krishnankutty et al. 2012). CDM is the process of collection, integration and validation of data in a clinical trial. High quality CDM ensures that data collected for analysis is accurate and complies with the protocol-specified requirements (Krishnankutty et al. 2012).

The main purpose of this chapter is to provide a synopsis of CDM for N-of-1 trials. For larger scale trials including multinational trials or trials evaluating investigational products for the purpose of registration of a new product, we refer the reader to the guidelines by the working group on data centers of the European

H. Senior (✉)
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: h.senior@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_8

Clinical Research Infrastructures Network (ECRIN), and websites of relevant country-specific regulatory authorities. Throughout the chapter, we provide references to the guidelines on the conduct of clinical trials produced by the International Conference on Harmonization (ICH) (International Conference on Harmonisation 2015), specifically, the guideline E6 which provides guidance on good clinical practice (GCP).

The chapter will describe the roles and responsibilities of members of the CDM team, and the various procedures including quality control required in CDM. Further, the rationale for and the procedure of trial registration is described at the end of the chapter.

Roles and Responsibilities of CDM Team Members

The size of a CDM team in a small study depends on the budget and the scope of the project. Often, research staff such as project managers may need to take on additional CDM roles.

The CDM team members are typically data managers, database programmers/designers, medical coders, clinical data coordinator, data entry associate and a quality control associate (Krishnankutty et al. 2012).

The investigators also have a critical role in CDM in ensuring data quality by collecting data that are accurate, complete and verifiable.

The data manager oversees the entire CDM process and liaises with the study researchers and project manager (also known as the clinical research coordinator). They are responsible for preparing the data management plan which describes the database design, data entry, data tracking, quality control measures, serious adverse event (SAE) reconciliation, discrepancy management, data extraction (including assisting the statistician in preparing data sets for analysis) and database locking. They approve all CDM activities (Krishnankutty et al. 2012; McFadden 2007).

The database programmer/designer performs case report form (CRF) annotation, creates the database, and programs and tests the data validation system. Further, they design data entry screens including access via the web for data entry (Krishnankutty et al. 2012; McFadden 2007).

The medical coder codes the data for medical history, medications, co-morbidities, adverse events, and concomitant medications (Krishnankutty et al. 2012). The clinical data coordinator is responsible for CRF design, CRF completion guides, a data validation plan (DVP) and discrepancy management (Krishnankutty et al. 2012).

Quality control associates check accuracy of data entry and conduct audits of the data for discrepancies (Krishnankutty et al. 2012).

Data entry associates track CRFs and perform data entry into the database (Krishnankutty et al. 2012).

Designing Case Report Forms (CRFs)

ICH-GCP states that "all clinical trial information should be recorded, handled, and stored in a way that allows its accurate reporting, interpretation, and verification" (ICH E6 Section 2) (International Conference on Harmonisation 2015). To comply in a trial there must be standardized operating procedures for data collection, processing and storage.

Case Report Forms (CRFs) are printed, optical or electronic documents designed to record all of the protocol-required information to be reported to the sponsor on each trial subject (ICH-GCP E6) (International Conference on Harmonisation 2015).

The CRFs must seek data that are specified in the protocol, and any data that are required by regulatory bodies.

The types of CRFs required in a clinical trial may include the following forms: screening/eligibility form, demographics, randomization form, medical history, physical examination and vital signs, concomitant/ concurrent medications, key efficacy endpoints of the trial and other efficacy data, adverse events, and laboratory tests. If using validated questionnaires in a CRF, the CRF should maintain the integrity of the questionnaire to maintain validity.

CRFs must be written and formatted to be user-friendly and self-explanatory (Krishnankutty et al. 2012). Most CRFs seek data that requires multiple-choice responses including yes/no responses, pre-coded tables, validated questionnaires or Likert scales, which allows data being collected across subjects to have the same terminology and to be easily combined (Liu et al. 2010). Generally, if free-text is being sought, it is kept to the minimum, with the exception of the recording of adverse events.

Thought also needs to be given to by whom and when data are to be collected, for example, if baseline data include both cardiology department data and medical imaging department data, then it may be better for separate CRFs for each department instead of requiring each department to complete a section of the same form (McFadden 2007).

Often the data collected are conditional on a previous question. For example, if a response is "Yes" for a medical history of diabetes, additional information may be sought through dependent questions on insulin use or diet, but only for those who responded with a yes. These series of questions are termed conditional data fields (Liu et al. 2010).

All CRFs should contain a header with the trial identifier and name, site number, and the subject identification number stated. They should also contain a section at the end of the form for the investigator/subject who completes the CRF to sign and date the form. The signature should match those authorized to complete CRFs according to the study signature log. CRFs should have version information in the footer along with page numbers (X of Y). The version number should be the most recent and be the version that has been ethically approved and listed on the version control log. It is also helpful to name the form in the header, for example Form A may be the screening or enrolment form, Form X may be the adverse event form etc.

Each question or item in a CRF needs to be numbered to facilitate communication regarding data queries between the CDM and investigators and make forms easier to follow for the user.

It is the responsibility of the investigator who completes the CRFs to ensure the accuracy, completeness, legibility, and timeliness of the data reported to the sponsor in the CRFs and in all required reports (ICH E6 Section 4) (International Conference on Harmonisation 2015). This includes ensuring the header and all sections of the CRF are complete, all adverse events and serious adverse events are recorded, and any data discrepancies are amended.

Pre-testing of study procedures, CRF completion and study measurements prior to the trial going "live" ensures that procedures are performed according to standard operating procedures and the forms are understandable, usable and have good flow (McFadden 2007). It allows any amendments and associated ethical approvals to be made before the first patient is enrolled.

With each CRF, the form should have an accompanying CRF completion guideline, often on the front page of the CRF, which provides details on the filling of the form. The CRF completion guidelines will also include clear definitions for specific variables, for example, a clear clinical definition of a standard drink, or moderate exercise. Diagnostic definitions of medical conditions must be clear, for example, one clinician may consider systolic blood pressure of >130 mmHg as hypertension, whereas another considers it to be >140 mmHg (Liu et al. 2010). Further, different diagnostic laboratories may have different criteria to define normal values for a laboratory test. The sponsor/researchers should obtain from each laboratory in a trial a list of normal values and ranges. Further, the units of measurement for a variable in a CRF should be consistent with the laboratory standards accepted by the major laboratories to avoid any unnecessary conversions. The completion guideline may contain calculations for the conversion between units.

Often CRFs are annotated, where on the CRF the variable names are written next to the spaces where the investigator writes in the data. Annotated CRFs provide a link between the database and the questions on the CRF allowing the CDM team and statistician to know where data is located on a database for each question in a specific CRF.

Design and Development of a Clinical Data Management System

To ensure quality CDM, the CDM team in collaboration with the project manager should develop a Data Management Plan (DMP), which describes the data management procedures including listing roles and responsibilities of personnel, a final set of CRFs, design of the database, data entry procedure, data query rules, query handling, quality assurance including audit trail checks, electronic data transfer rules, database backup and recovery, archiving, database security, procedures for database locking and unlocking, and reports (Ohmann et al. 2011).

The development of CDM should begin in the early stages of a study. The database is created on the basis of the protocol-derived CRFs. The types of data including number of decimal points and units should be clear from the CRFs when developing the database.

Databases have to be designed to allow the linking of multiple database records for an individual subject, by including a unique subject number for each record (McFadden 2007). On the database, subjects must be de-identified, where each subject is given a unique subject number that serves as the subject identifier within the database (Liu et al. 2010).

When developing the database, the designer has to define the database to comply with the study objectives, intervals, visits, users with different levels of access, different sites and subjects.

Depending on the requirements of the project and the study budget, there is a range of software available to create a database from MS excel, MS access, open source software (TrialDB, OpenClinica, openCDMS, PhOSCo), to specialized software (Oracle clinical, Redcap, Clintrial, eClinical suite) among others. As a rule, these software packages are designed to comply with regulatory requirements for conducting clinical trials, but it is the responsibility of the sponsors to confirm this assumption.

If the purpose of a study is to provide evidence for regulatory approvals of new medications or new indications, the database must comply with ICH-GCP and country-specific regulations, and we recommend the employment of an experienced clinical data manager. For these studies, it is imperative to ensure that the software allows an audit trail of all activity in the database (Krishnankutty et al. 2012).

Security of the database is paramount, and most software allows the data manager to allocate access of users to only the parts of the database to which they are required to access according to roles and responsibilities, thereby preventing any unauthorized access or changes to the database. The data manager must maintain a list of individuals who are authorized to have access and make changes to the data. Further, the database must be constructed in such a way to ensure the safeguarding of any trial blinding.

If a database user enters data, most software provides an audit trail by recording the changes made, the user's details, and the date and time of the change (Krishnankutty et al. 2012). Further, the database must have a system of automatic backups to ensure data preservation.

With all CDM, it is important to develop a system to maintain quality control of the data during all stages of the trial. ICH-GCP states that "Quality control should be applied to each stage of data handling to ensure that all data are reliable and have been processed correctly" (ICH-GCP E6 Section 5) (International Conference on Harmonisation 2015).

Subject Enrolment and Data Entry

After subjects have provided informed consent, and prior to the collection of data, subjects are enrolled into a clinical trial using an "enrolment form". The enrolment form is used to screen subjects for eligibility into the trial using the inclusion and exclusion criteria as stated in the ethically approved protocol. If a subject is not eligible, the reasons for the ineligibility must be retained to allow reporting of the trial according to the CONSORT statements for reporting trials (Schulz et al. 2010). If eligible, the subject is randomized into either the intervention or control arm of the trial and provided with a unique subject number. At this time, the contact details of the subject should be collected on a separate contact details form, along with details of a next of kin who does not live with the subject, to allow follow-up. The subject number must be recorded on this form, and the form stored separately to the CRFs or any other documents containing patient medical information or data. CDM may be involved in randomizing subjects after enrolment by providing a randomization service either electronically or via telephone/fax using randomization lists or programs.

An additional form required for all trials is a 'serious adverse event (SAE) report form'. Refer to the chapter on SAEs (Chap 10) for more information on the type of data to collect and the processes for the evaluation and reporting of SAEs.

Data in clinical trials can be collected using either a paper CRF (p-CRF) or an electronic CRF (e-CRF). For some N-of-1 trials, data are collected on p-CRFs followed by data entry into the database by research staff. The p-CRFs are completed by the investigator, and sometimes the subject (which is termed *patient reported outcomes*), according to the completion guidelines (Krishnankutty et al. 2012). In contrast, e-CRFs allow data to be entered directly into the database through a computer at the study site.

The e-CRF method of data entry, also called remote data entry, has advantages over p-CRFs in that data are entered sooner, there is no risk of loss of forms, data discrepancies can be raised and resolved more quickly, and the chance of transcription error is less (Krishnankutty et al. 2012). However sponsors need to consider if potential sites have access to computers with high-speed internet access and the cost of using e-CRFs (Liu et al. 2010). Mobile phone technology is also developing as a tool for remote data entry. For smaller studies as often occurs with N-of-1 trials, data collection on p-CRFs followed by entry to a centralized study database is most likely to be the most cost-effective method.

Many fields in a CRF for a clinical trial require the abstraction of data from a source document. Source documents contain "source data" which are defined as "all information in original records and certified copies of original records of clinical findings, observations, or other activities in a clinical trial necessary for the reconstruction and evaluation of the trial" (ICH-GCP E6 Section 1) (International Conference on Harmonisation 2015). The major exception where CRF recorded data will not have a source are those CRFs that contain data completed by the subjects, and the data have not been transcribed by the investigators onto separate CRFs (Liu et al. 2010).

A source document can include original documents, data and records (e.g., hospital records, clinical and office charts, laboratory notes, memoranda, subjects' diaries or evaluation checklists, pharmacy dispensing records, recorded data from automated instruments, copies or transcriptions certified after verification as being accurate copies, microfiches, photographic negatives, microfilm or magnetic media, X-rays, subject files, and records kept at the pharmacy, at the laboratories and at medico-technical departments involved in the clinical trial) (ICH-GCP E6 Section 1) (International Conference on Harmonisation 2015).

On the CRFs, subjects must be de-identified. Each subject is given a unique subject number that serves as the subject identifier within the database (Liu et al. 2010). It is a sound idea to also record the subject's initials on the CRFs along with the unique subject number. Identifiers that should be separate from the CRFs and database are names, addresses, email addresses, contact phone numbers, social security numbers or equivalent, medical record numbers, and photos (Liu et al. 2010). Specifically, this is any information where there is a reasonable basis to believe the information can identify the individual.

The requirement of de-identification of subject's data is re-iterated in ICH-GCP, which states that "the confidentiality of records that could identify subjects should be protected, respecting the privacy and confidentiality rules in accordance with the applicable regulatory requirements"(ICH E6 Section 2) (International Conference on Harmonisation 2015).

The subject's name and initials will appear on a subject log along with the unique subject number and this log is stored separately to the CRFs and the database. The subject log and a contact details form should be available for the verification of source documents and subject follow-up (Liu et al. 2010).

After completion of a CRF, if an investigator wishes to make a correction or change to a CRF, they should put a line through the original data without obscuring the original entry, write the new data alongside, date and initial the change (and the initials should match those in the signature log), and provide an explanation for the change. This applies to both written and electronic changes or corrections (ICH E6 Section 4) (International Conference on Harmonisation 2015). Most CDM computer software automatically records any changes made on electronic CRFs including the identity of the investigator, the original data, and date and time of the change.

Submission of study data to the CDM team may be by mail, courier, fax or electronically (Liu et al. 2010). Timely submission is critical to ensure deadlines for data entry are met (Liu et al. 2010) and data validation can occur. SAE report forms are usually required within 24 h of the investigator becoming aware of an SAE. CRFs can be submitted when completed or at specified intervals depending on the process for verification of source documents at the site (Liu et al. 2010).

To minimize the risk of transcription errors, data are often entered from p-CRFs into e-CRFs using double data entry, often by two operators separately. Any discrepancy is checked and resolved, leading to a cleaner database (Krishnankutty et al. 2012). With single data entry, the data operator must double-check that the data entered matches exactly the data on the p-CRF, and retain the p-CRFs in the subject folder.

According to ICH-GCP, every clinical trial must have a designated staff member who will monitor trial activities, called a clinical research associate (CRA) or a study monitor. One role of a CRA is monitoring data entry into a CRF to ensure completeness of data. The CDM team tracks CRFs to ensure data is collected at the right assessment point, to check for missing pages on CRFs or illegible data, to raise data discrepancies to the investigator seeking resolution, and to ensure data completeness, timeliness and accuracy.

An important role of the CRA is to conduct a site visit to ensure that data recorded on a CRF which has been derived from a source document is consistent with the source document, and if inconsistent, the discrepancy is explained by the investigator (ICH-GCP E6 Section 4) (International Conference on Harmonisation 2015). If there are any data on a CRF where there is a discrepancy between the data entry in the CRF and the source document, the CRF needs to be corrected to match the data in the source document or an explanation provided as to why the CRF is correct.

Data Validation

To ensure the quality of data in a database, data validation occurs throughout data collection. Data validation is the process of testing the validity of the data against protocol specifications (Krishnankutty et al. 2012). The project manager must provide the CDM team with an edit specifications list which describes which data are to be checked and queried, and this is programmed into the database. If the investigators are not using a database with edit checking capabilities, they will have to conduct edit checks manually prior to any data entry of a CRF.

Edit check programs test each variable with a logic condition specific for the variable. All edit check programs are tested with dummy data before the database goes live to ensure they are working (Krishnankutty et al. 2012). Any discrepancy that occurs between the logic condition and the variable will raise a data query. The types of discrepancies may include missing data that must be entered, data that is not consistent with other data recorded on the forms, data out of range (range checks), and protocol deviations (Krishnankutty et al. 2012; Liu et al. 2010).

Examples of edit checks include checking if eligibility criteria are met prior to randomization, and checking that variables like height and weight are within a certain range. Edit checks need to be logical. It may be that a person may be taller than the range, as the subject is more than the pre-specified two standard deviations above the population mean height. This discrepancy would be resolved by responding that the height is correct.

With e-CRFs, many of the edit checks can occur immediately, raising data queries as data are entered into the database. Other data queries will be raised as data validation processes are conducted at regular intervals, and the investigator will resolve any queries after logging into the system (Krishnankutty et al. 2012). Further, data queries may be logged on a study report or data clarification form derived from the database which the project manager, CDM team and investigator can access to ensure the investigator resolves queries quickly.

The CDM team are responsible for data quality checks to identify missing, illogical and inconsistent data both automatically in batches, and manually. Manual checks will review the CRFs for any missed discrepancies such as laboratory data or medical records recorded on a CRF suggesting an adverse event has occurred, but this event is not recorded on the serious adverse event or adverse event forms.

Management of Data Discrepancies and Study Reports

Query resolution requires the investigator to review the discrepancies, investigate the reason for the discrepancy, and resolve the discrepancies based on documents (Krishnankutty et al. 2012). Alternatively, the discrepancy may be declared by the investigator to be unresolvable. There are two purposes of this process, the cleaning of the data and the gathering of evidence for any "irresolvable" discrepancies in the database (Krishnankutty et al. 2012).

When a data query is raised, the investigator will enter the correct data or explain the reasons behind the discrepancy (Krishnankutty et al. 2012). An important role of the project manager and/or CDM team is to regularly review all discrepancies to ensure they are resolved (Krishnankutty et al. 2012). Once a discrepancy is resolved, it no longer will appear on a data clarification report (Krishnankutty et al. 2012).

To assist with the management of data discrepancies and overall project management by the project manager, the CDM team and the investigators, the study database can be utilized to generate study reports. The most common study reports are patient accrual, missing data, missing forms, forms entered per month, adverse events, scheduled visits, and inconsistent or invalid data.

Medical Coding

Often data are collected in clinical trials that require medical coding, which is a process of classifying disease sub-types, medical procedures, medical terminologies and medications. In large scale trials, the coding will be entered directly into the database by a qualified medical coder. However, in smaller studies with limited resources, the research or CDM team will have to assign a person with appropriate medical knowledge to conduct the medical coding using a medical dictionary and knowledge of the hierarchy of classifications in medical coding to allow coding within proper classes. The types of data that may require medical coding to allow counts in statistical analysis include disease sub-types, medical procedures, and medications.

Common dictionaries for coding include the Medical Dictionary for Regulatory Activities (MedDRA) for coding of adverse events, medical history, medical terms; the WHO Adverse Reactions Terminology (WHOART) for the coding of adverse events, and the WHO Drug Dictionary Enhanced (WHO-DDE) (Krishnankutty et al. 2012; Liu et al. 2010). The dictionaries to be used for coding must be specified in the protocol.

Medical coding classifies medical terminology on CRFs to achieve data consistency and subject safety, thereby ensuring that even though the same adverse event may be described differently by different investigators, or investigators record different drug trade names for the same medication, the data is coded correctly according to a uniform standard (Krishnankutty et al. 2012).

Database Close-Out, Data Lock and Archiving

Prior to database close-out, a final data quality check and validation is made to ensure there are no discrepancies remaining that have not been assessed to be "irresolvable". All study activities are completed including medical coding and data cleaning procedures, and external data is reconciled (Ohmann et al. 2011). Datasets required by the study statistician are finalized (Krishnankutty et al. 2012). After data-lock, no data can be changed in the database. Upon approval of the steering committee, the database is locked and the data is extracted for data analysis (Krishnankutty et al. 2012).

After data extraction, the database is archived. The data may be archived along with essential documents on CDs/DVDs. The length of archiving of essential documents can range between 2 to 15 years. If there is a marketing application, ICH-GCP requires archiving for a minimum of 2 years after the approval of application, or if no marketing application is made, records must be archived for 2 years after the end of the trial (ICH-GCP E6 Section 4) (International Conference on Harmonisation 2015). Sponsors and researchers must be aware of the regulations and guidelines on retention of records for their specific local and national regulatory bodies, for example, in trials involving children in Australia the period of retention can be 28 years. If documents are archived off-site due to space constraints, then a record of the location of the records must be kept by the sponsor or researcher (Liu et al. 2010).

Registering a Clinical Trial in a Trial Register

From July 2005, the International Committee of Medical Journal Editors (ICMJE) adopted a policy (De Angelis et al. 2005) that all member journals will only consider a trial for publication if it has previously undergone registration in a trial registry.

N-of-1 trials are not exempt from this position of the ICMJE. A clinical trial is defined by the ICMJE as "any research project that prospectively assigns human subjects to intervention or comparison groups to study the cause-and-effect relationship between a medical intervention and a health outcome" (De Angelis et al. 2005).

The purpose of a trial registry and the policy statement of the ICMJE is the substantial under-reporting of the findings of clinical trials. Under-reporting of clinical trials potentially leads to a bias of the overall knowledge of the effect of a medical intervention, leading to over-estimates of benefit and under-estimates of harm

(McGauran et al. 2010; Chalmers et al. 2013) Under-reporting of trials occurs both within commercially sponsored trials and academic trials (Chalmers et al. 2013).

In clinical practice, evidence for interventions that will have an impact in many cases does not arise from a single trial, but from the collective body of evidence, as assessed by a systematic review and meta-analysis. As such, selective reporting misrepresents the true effectiveness of a medical intervention and negatively impacts on both clinical guidelines and practice. Registering trials prior to the conduct of a trial, places the awareness of the trial in the public domain.

An additional important ethical consideration is that subjects who volunteer for a clinical trial do so because they assume that they are advancing medical knowledge. This places an obligation on the trial sponsor and investigators not to betray this trust (Chalmers et al. 2013; De Angelis et al. 2005) This responsibility is clearly stated in the Helsinki declaration that states that "Every clinical trial must be registered in a publicly accessible database before recruitment of the first subject"(World Medical Association 2008). In some countries, clinical trials must be registered to comply with policies from funding bodies or legislation (including the European Commission, and the US Food and Drug Administration).

Clinical trials can be registered in ICMJE approved registries (International Committee of Medical Journal 2015) or within the WHO registry network (World Health Organization 2014). A trial only needs to be registered once, although some investigators may register a trial in more than one country depending on specific funders' or countries' policies and regulations.

Conclusion

At the end of the hard work of seeking funding, preparing and conducting a clinical trial to answer an important research question, ideally researchers will have produced an accurate and complete database for analysis. The phenomenon of 'garbage in, garbage out' (GIGO) applies to clinical trials. Databases that are incomplete and/or inaccurate increase the risk of biased findings. To avoid the GIGO phenomenon in N-of-1 trials, it is imperative to engage a clinical data management system that adopts well-designed and user-friendly CRFs and databases, a data validation and CRF tracking system, an audit trail, and clinical monitoring.

References

Chalmers I, Glasziou P, Godlee F (2013) All trials must be registered and the results published. BMJ 346:1–2

De Angelis CD, Drazen JM, Frizelle FA, Haug C, Hoey J, Horton R, Kotzin S, Laine C, Marusic A, Overbeke AJ, Schroeder TV, Sox HC, Van Der Weyden MB, International Committee of Medical Journal, Editors (2005) Is this clinical trial fully registered? a statement from the international committee of medical journal editors. Lancet 365:1827–1829

International Committee of Medical Journal Editors (2015) *ICMJE clinical trials registration* [Online]. Available: http://www.icmje.org/about-icmje/faqs/clinical-trials-registration/

International Conference on Harmonisation (2015) *ICH official web site : ICH* [Online]. Available: http://www.ich.org/home.html

Krishnankutty B, Bellary S, Kumar NB, Moodahadu LS (2012) Data management in clinical research: an overview. Indian J Pharmacol 44:168–172

Liu MB, Davis K, Liu MB, Duke Clinical Research Institute (2010) A clinical trials manual from the duke clinical research institute : lessons from a horse named Jim. Wiley-Blackwell, Chichester/UK/Hoboken

Mcfadden E (2007) Management of data in clinical trials. Wiley-Interscience, Hoboken

Mcgauran N, Wieseler B, Kreis J, Schuler YB, Kolsch H, Kaiser T (2010) Reporting bias in medical research – a narrative review. Trials 11:37

Ohmann C, Kuchinke W, Canham S, Lauritsen J, Salas N, Schade-Brittinger C, Wittenberg M, Mcpherson G, Mccourt J, Gueyffier F, Lorimer A, Torres F, CTR EWGD (2011) Standard requirements for GCP-compliant data management in multinational clinical trials. Trials 12:85

Schulz KF, Altman DG, Mohe D, Group, C (2010) Consort 2010 statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332

World Health Organization (2014) *The WHO registry network* [Online]. Available: http://www.who.int/ictrp/network/en/

World Medical Association (2008) *Ethical principles for medical research involving human subjects* [Online]. Available: http://www.wma.net/en/30publications/10policies/b3/17c.pdf

Chapter 9
Individual Reporting of N-of-1 Trials to Patients and Clinicians

Michael Yelland

Abstract This chapter offers a very practical account of the reporting of N-of-1 trials to patients and clinicians, using trials for chronic pain conditions as models which may be applied to many other forms of N-of-1 trials. It draws from the author's experience in managing N-of-1 trials comparing celecoxib with extended release paracetamol for chronic pain and osteoarthritis and comparing gabapentin with placebo for chronic neuropathic pain· Reporting the results of N-of-1 trials to patients and health care professionals requires considerable planning to make reports user-friendly and an efficient tool for clinical decision making. Decisions need to be made about key elements of the report, how to order them with the most important summary elements first followed by detailed results, and how to set thresholds for clinically important changes. The inclusion of tables and graphs in reports should improve readability. An example of an individual report is provided.

Keywords N-of-1 trials • Reporting • Results • Threshold • Quantitative analysis • Qualitative analysis • Clinically important change • Clinical response • Threshold • Minimal detectable change • Adverse event • Overall response

Aim of Reporting

The reporting processes of N-of-1 trials can be viewed as analogous to reporting on other investigations in medicine, for example radiology reports. There is a need to describe what was done, the results that arose from the investigation and finally the interpretation of the results in the form of a summary or conclusions.

M. Yelland MBBS, Ph.D., FRACGP, FAFMM. Grad Dip Musculoskeletal Med (✉)
School of Medicine, Griffith University and Menzies Health Institute,
Gold Coast, QLD, Australia
e-mail: m.yelland@griffith.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_9

What Should Be Reported?

Decisions must be made about the elements of the trial that are essential to report. Key elements of the report may include:

- Patient details
- Description of the trial – medications compared; order of medication periods; marker joint/region; date of report
- Conclusion/summary of overall response
- Summary of outcomes used to determine the overall response
- Use of other medications during the trial
- Success of blinding
- Detailed results of the individual outcome measures, including graphs and/or tables of relevant data points.

The results we reported were a mix of quantitative outcomes and qualitative outcomes. See Fig. 9.1 at the end of this chapter for an example of the report that we generated. The quantitative outcomes included mean scores of numerical rating scales with ranges of zero to ten on which the severity of symptoms, functional loss (Yelland et al. 2007; Yelland et al. 2009) or sleep disturbance (Yelland et al. 2009) were rated. In arriving at these means, we omitted the first week of data from each two week period to negate any carry-over effects from the preceding period.

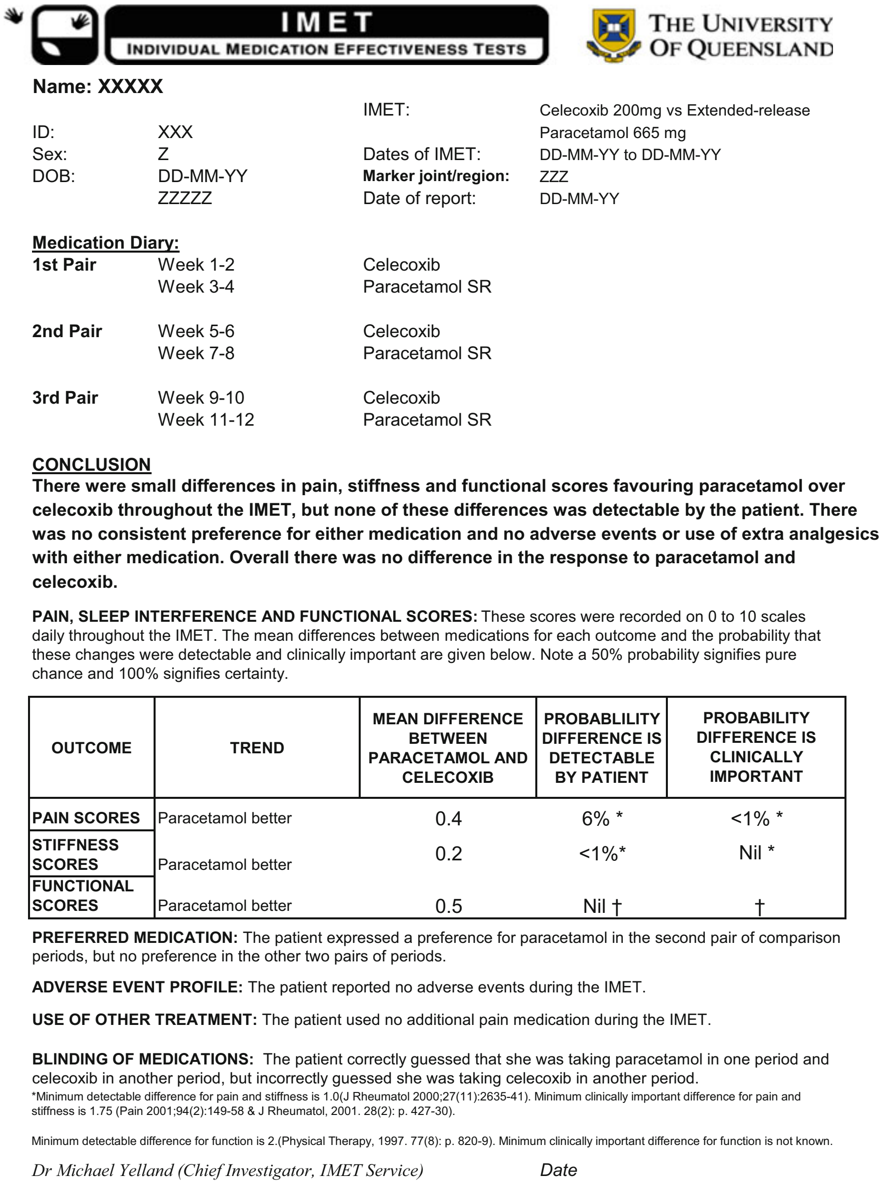
The qualitative outcomes included medication preference between the current period and the preceding period and a summary of adverse events during treatment periods. Medication preference was recorded as a preference for one of the two medications or no preference. Adverse events were listed and tallied for each period.

How Should the Data Be Presented?

Reporting the results of N-of-1 trials should be in a format that suits the needs of the ‘consumers’ of the trial service, namely the patient and their health care professional. Their needs may differ, with some wishing to just read the ‘headlines’ or conclusions, and others wishing to read the fine details that underlie these conclusions. For this reason we chose to put the conclusions very early in the report, directly after the description of the order of medication periods within the trial.

The raw data from treatment periods can be reported in graphical, tabular or descriptive formats depending on its nature. Data from washout periods where a carry-over effect may apply can be omitted. For quantitative data collected on a daily or weekly basis, we used graphs of scores over time. The means of scores for each treatment period were presented in a table to allow easy comparison between treatment periods (Fig. 9.1).

The qualitative data on medication preferences and description of adverse events were presented in tabular format.



IMET
INDIVIDUAL MEDICATION EFFECTIVENESS TESTS

THE UNIVERSITY OF QUEENSLAND

Name: XXXXX

		IMET:	Celecoxib 200mg vs Extended-release Paracetamol 665 mg
ID:	XXX		
Sex:	Z	Dates of IMET:	DD-MM-YY to DD-MM-YY
DOB:	DD-MM-YY	**Marker joint/region:**	ZZZ
	ZZZZZ	Date of report:	DD-MM-YY

Medication Diary:

1st Pair	Week 1-2	Celecoxib
	Week 3-4	Paracetamol SR
2nd Pair	Week 5-6	Celecoxib
	Week 7-8	Paracetamol SR
3rd Pair	Week 9-10	Celecoxib
	Week 11-12	Paracetamol SR

CONCLUSION

There were small differences in pain, stiffness and functional scores favouring paracetamol over celecoxib throughout the IMET, but none of these differences was detectable by the patient. There was no consistent preference for either medication and no adverse events or use of extra analgesics with either medication. Overall there was no difference in the response to paracetamol and celecoxib.

PAIN, SLEEP INTERFERENCE AND FUNCTIONAL SCORES: These scores were recorded on 0 to 10 scales daily throughout the IMET. The mean differences between medications for each outcome and the probability that these changes were detectable and clinically important are given below. Note a 50% probability signifies pure chance and 100% signifies certainty.

OUTCOME	TREND	MEAN DIFFERENCE BETWEEN PARACETAMOL AND CELECOXIB	PROBABLILITY DIFFERENCE IS DETECTABLE BY PATIENT	PROBABILITY DIFFERENCE IS CLINICALLY IMPORTANT
PAIN SCORES	Paracetamol better	0.4	6% *	<1% *
STIFFNESS SCORES	Paracetamol better	0.2	<1%*	Nil *
FUNCTIONAL SCORES	Paracetamol better	0.5	Nil †	†

PREFERRED MEDICATION: The patient expressed a preference for paracetamol in the second pair of comparison periods, but no preference in the other two pairs of periods.

ADVERSE EVENT PROFILE: The patient reported no adverse events during the IMET.

USE OF OTHER TREATMENT: The patient used no additional pain medication during the IMET.

BLINDING OF MEDICATIONS: The patient correctly guessed that she was taking paracetamol in one period and celecoxib in another period, but incorrectly guessed she was taking celecoxib in another period.

*Minimum detectable difference for pain and stiffness is 1.0(J Rheumatol 2000;27(11):2635-41). Minimum clinically important difference for pain and stiffness is 1.75 (Pain 2001;94(2):149-58 & J Rheumatol, 2001. 28(2): p. 427-30).

Minimum detectable difference for function is 2.(Physical Therapy, 1997. 77(8): p. 820-9). Minimum clinically important difference for function is not known.

Dr Michael Yelland (Chief Investigator, IMET Service) Date

Fig. 9.1 Deidentified report of a patient who completed an individual medication effectiveness test or N-of-1 trial on celecoxib versus sustained-release paracetamol for osteoarthritis.

DETAILED RESULTS

PAIN, STIFFNESS AND FUNCTIONAL SCORES

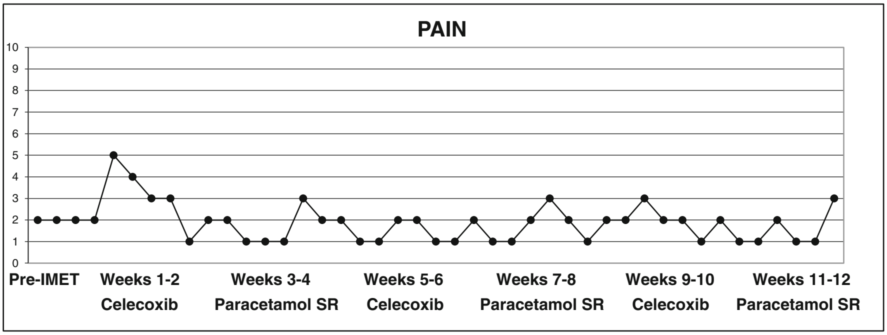
Week	Medication	Medication Guess	Average pain score*	Average stiffness score^	Average functional score~
Pre-IMET	*Celecoxib*	*N/A*	*2*	*5*	*4.40*
Week 1-2	Celecoxib	Unsure	3.00	2.71	2.75
Week 3-4	Paracetamol SR	Unsure	1.57	1.86	2.40
Week 5-6	Celecoxib	Unsure	1.57	1.71	2.20
Week 7-8	Paracetamol SR	Celecoxib	1.71	2.00	2.25
Week 9-10	Celecoxib	Celecoxib	1.86	1.71	2.40
Week 11-12	Paracetamol SR	Paracetamol SR	1.57	1.71	1.80

*0 = No pain, 10 = Extreme pain
^0 = No stiffness, 10 = Extreme stiffness
~0 = Unable to complete activity, 10 = Able to perform activity at pre-arthritis level

PREFERRED MEDICATION AND ADVERSE EVENTS

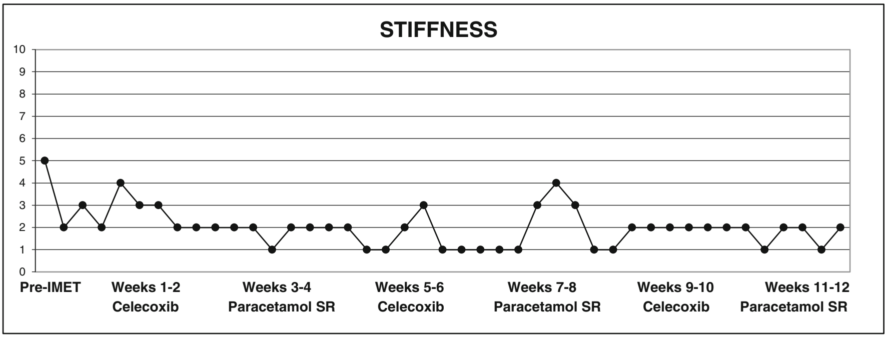
Week	Actual Medication	Preferred medication	No. of adverse events	Details of adverse events
Week 1-2	Celecoxib	Unsure	0	N/A
Week 3-4	Paracetamol SR		0	N/A
Week 5-6	Celecoxib	Paracetamol	0	N/A
Week 7-8	Paracetamol SR		0	N/A
Week 9-10	Celecoxib	Both the same	0	N/A
Week 11-12	Paracetamol SR		0	N/A

GRAPHS

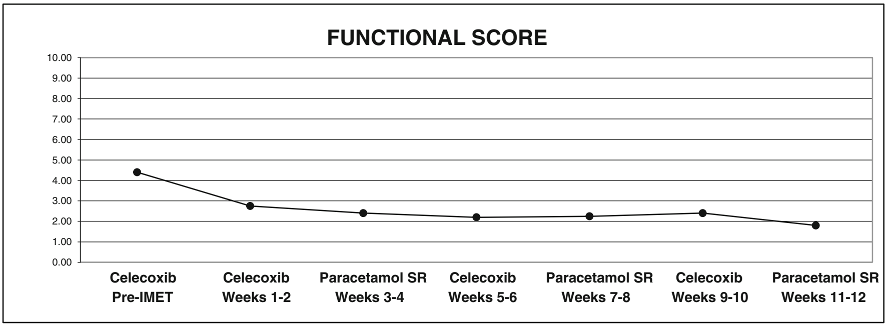


*0 = No pain, 10 = Extreme pain

Fig. 9.1 (continued)



^0 = No stiffness, 10 = Extreme stiffness



~0 = Able to perform activity at pre-arthritis level, 10 = Unable to perform activity

NOTE A HIGHER SCORE INDICATES MORE DIFFICULTY PERFORMING FUNCTION

Fig. 9.1 (continued)

Defining Response for Each Outcome

Defining what constitutes a response for each outcome can present some challenges. This is in part because of the different ways outcomes are recorded and in part because the threshold for response varies from individual to individual. With quantitative outcomes, the traditional way of defining differences in population trials is to test for statistical significance of the difference between the mean change over time for two groups of participants. In single patient datasets, the difficulties with using inferential statistics with a threshold of significance of 0.05 % or 5 % have been discussed elsewhere in this book in the section on analysis. Bayesian methods are more appropriate for calculating and expressing differences in response to treatment within the individual (Zucker et al. 1997). In Bayesian statistics results are expressed as the probability that a nominated trend is present, e.g. there is an 87 %

probability that pain scores on drug A are lower than on drug B. This is more informative than reporting whether or not the difference in outcomes met a predetermined threshold of statistical significance.

Nonetheless, it is desirable to have some method of setting a threshold for a difference in response between treatments over the course of each trial. This will allow comparisons between individuals in a series of N-of-1 trials. For this some guidance may be found in the literature on minimum clinically important differences (MCID), minimum clinically important change (MCIC) and minimum detectable change (MDC). These will have been derived for some outcomes from mean results in population based clinical trials and so may not represent what an individual patient regards as important or detectable. This could be determined prospectively in consultation with the patient with a question such as "*What is the minimum percentage improvement in your (insert outcome) that would make this treatment worthwhile?*" (Yelland and Schluter 2006).

The MCID which is defined as "the smallest difference in score in the domain of interest which patients perceive as beneficial and would mandate, in the absence of troublesome side effects and excessive cost, a change in the patient's management"(Jaeschke et al. 1989). This can be calculated by the differences between groups in a clinical trial. In the case of the MCIC, it is calculated by the differences within individuals in a clinical trial (Bombardier et al. 2001). Many methods of estimating these values exist (Copay et al. 2007). The commonest are 'anchor-based' methods that compare change scores in key outcomes over time with an anchor of the retrospective assessment of global response by the patient. For example, the MCIC for pain may be the mean change in pain scores of those who regard themselves as 'much better'. Alternatively it may be the difference in mean change scores between those who rate their improvement as 'much better' and 'very much better' from the remainder who did not do as well as this.

The other way of estimating clinically important changes is by using distribution methods.(Copay et al. 2007) One distribution method is based on the effect size, which is calculated by subtracting the mean of the scores at baseline from the mean of the scores at the follow-up point and then dividing this difference by the standard deviation at baseline.

An effect size of 0.5, described as 'moderate,' seems to correlate best with MCIC calculated by anchor based methods. However standard deviations can vary considerably from sample to sample, so some prefer other distribution-based methods that use the standard error of the mean (SEM) to calculate what is called the 'minimum detectable change'(MDC). The SEM provides a measure of within-person change that is less dependent on a specific sample because it incorporates both the standard deviation and the reliability. The MDC is equivalent to one SEM and represents the minimum change that is reliably detected by patients (Copay et al. 2007).

In our chronic pain trials, we reported both the MCID and the MDC for pain, but used the MDC to define a response. Using the published MDC of 1.0 for pain scores (Dworkin et al. 2008), a definite response was defined as an adjusted mean absolute difference ≥1.0, a probable response as a difference of ≥0.5 but <1.0, and differences of less than 0.5 as no response.

For categorical outcomes a different method of defining the threshold for a response is needed. For example, the medication preference within each comparison has three possible responses – a preference for drug A, a preference for drug B or no preference. We defined a definite response a preference for one medication in all 3 comparisons, a probable response as a preference in 2 comparisons and no response as a preference in 1 comparison or none.

Yet another method is necessary for undesirable outcomes, such as adverse events. While it is possible to develop elaborate scoring systems for adverse events that take into account the number, frequency and severity of adverse events, for the sake of simplicity we chose to define responses based on the numbers of events in each treatment period. Here, a definite response was defined as fewer events on one medication in all three comparisons, a probable response as fewer events in two comparisons and no response as fewer events in no comparisons or one comparison. The weakness of this approach is that it does not recognize the difference in 'bothersomeness' of adverse events. It may be more informative to get the patient to classify their adverse events by their bothersomeness into 'minor', 'moderate' or 'major' and create an algorithm that incorporates this when defining a response.

Defining an Overall Response

A statement on the overall response conveniently provides a single result that summarises the trial outcomes for the patient and clinician and is also useful in reporting on series of N-of-1 trials in scientific papers. However it should not necessarily be the one measure that determines future treatment decisions. Defining an overall response requires the aggregation and integration of the results from several outcomes into a single result. This is not an easy task as it is necessary to make a judgment about the relative value of each outcome. This may be at odds with the relative value of each outcome for individual patients. Symptom relief may be the most important outcome for a highly symptomatic patient, whilst absence of adverse events may be more important to one who has suffered a lot of adverse events in the past. We dealt with this dilemma in the celecoxib-paracetamol trials by creating an aggregate response variable, composed from the five outcomes weighted equally (Yelland et al. 2007). Each outcome was arbitrarily defined on a 5-point scale from −2 favouring celecoxib to +2 favouring paracetamol. An individual with aggregate response absolute value ≥6 was considered a definite responder, a value ≥3 but <6 was considered a probable responder, and a value <3 was considered a non-responder.

Equal weightings were assigned to each outcome here in the belief that it was impossible to have a valid system of weighting each outcome. However there is now an emerging science of discrete choice experimentation that allows determination of an average value patients place on different attributes when making health related decisions. This could conceivably be used to determine the relative value of outcomes in a series of patients undertaking N-of-1 trials. These relative values could be used to design more representative methods of creating aggregate scoring systems (Ryan 2004). Alternatively, it would be better for the individual to assign a relative value to each outcome on a score of 0 to 10 and factor this into the calculation of the aggregate score.

Reporting to the Patient and the Health Care Professional

As the essence of N-of-1 trials is to inform clinical decision-making, timely reporting of results to the patient and health care professional is essential. This involves a responsive system for analyzing results and preparing and finalizing reports before sending them out. We undertook to send out reports within two weeks of receiving the final results from the patient. They were faxed and posted in the period from 2000 to 2005, but since then more efficient and secure electronic methods of transmission have become available.

A decision should be made about what treatment, if any, is continued whilst awaiting the results. In the gabapentin-placebo trial (Yelland et al. 2009), we continued the supply of gabapentin for the patient until they had discussed the result with their doctor. After this discussion we sent a brief questionnaire to both the patient and doctor enquiring about medication decisions arising from the trial. Patients were subsequently followed up by telephone at 3, 6 and 12 months to look at concordance with these treatment decisions.

Conclusion

In summary, reporting the results of N-of-1 trials to patients and health care professionals requires considerable planning to make reports user-friendly and an efficient tool for clinical decision-making. Decisions need to be made about key elements of the report, how to order them with the most important summary elements first followed by detailed results, and how to set thresholds for clinically important changes. The inclusion of tables and graphs in reports should improve readability. Transmission of reports to patients and their health care professionals should be done very soon after completion of the trial when the results are most useful for clinical decision–making.

References

Bombardier C, Hayden J, Beaton DE (2001) Minimal clinically important difference low back pain: outcome measures. J Rheumatol 28:431–438

Copay AG, Subach BR, Glassman SD, Polly DW, Schuler TC (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 7(5):541–546

Dworkin RH, Turk DC, Wyrwich KW, Beaton D, Cleeland CS, Farrar JT et al (2008) Interpreting the clinical importance of treatment outcomes in chronic pain clinical trials: IMMPACT recommendations. J Pain 9:105–121

Jaeschke R, Singer J, Guyatt GH (1989) Measurement of health status. Ascertaining the minimally clinically important difference. Control Clin Trials 10:407–415

Ryan M (2004) Discrete choice experiments in health care. BMJ 328:360–361

Yelland M, Schluter P (2006) Defining worthwhile and desired responses to treatment of chronic low back pain. Pain Med 7(1):38–45

Yelland MJ, Nikles CJ, McNairn N, Del Mar CB, Schluter PJ, Brown RM (2007) Celecoxib compared with sustained-release paracetamol for osteoarthritis: a series of N-of-1 trials. Rheumatology (Oxford) 46(1):135–140

Yelland MJ, Poulos CJ, Pillans PI, Bashford GM, Nikles CJ, Sturtevant JM (2009) N-of-1 randomized trials to assess the efficacy of gabapentin for chronic neuropathic pain. Pain Med 10(4):754–761

Zucker DR, Schmid CH, McIntosh MW, D'Agostino RB, Selker HP, Lau J (1997) Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. J Clin Epidemiol 50(4):401–410

Chapter 10
Assessing and Reporting Adverse Events

Hugh Senior

Abstract The safety of subjects who volunteer to participate in clinical trials is paramount. The (ICH-Good Clinical Practice) guidelines assert that ‘the rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society’. This chapter describes the internationally accepted standard of the ICH-GCP guidelines. It introduces important clinical research terminology, and provides definitions of various types of adverse events, describes the roles and responsibilities of investigators and sponsors, and the processes needed to promote safety through the assessment, recording, evaluating and reporting of adverse events during the design and conduct of clinical trials.

Keywords Adverse events • Adverse drug reactions • ICH-GCP • SUSAR • Ethics • Data monitoring • Case report form • Causality • Seriousness • Expectedness

Introduction

To ensure mutual acceptance of clinical data, representatives of industry, academia and Governmental and non-Governmental health organizations have developed and agreed on a set of guidelines for the conduct of clinical trials. Three regions were represented in discussions, namely, the European Union, the United States and Japan, as well as the World Health Organization, Canada, Australia, and the Nordic Countries. The guidelines are known as the ICH-GCP (ICH-Good Clinical Practice) guidelines. ICH-GCP is the recognized standard internationally for the conduct of clinical trials, including N-of-1 trials.

The guidelines provide the minimal ethical and scientific standard regarding the design, conduct, and reporting of clinical trials involving human participants. The guidelines can be found on the official ICH website (International Conference on Harmonisation 2015).

H. Senior (✉)
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: h.senior@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_10

It is paramount that all investigators involved in clinical trials are trained in and are fully aware of the guidelines. An essential guideline is the efficacy guidelines, denoted by the letter ‘E’, especially ‘E6’, which is the guideline addressing Good Clinical Practice (GCP), the last version of which was in May 1996. GCP is defined as ‘a standard for the design, conduct, performance, monitoring, auditing, recording, analyses, and reporting of clinical trials that provides assurance that the data and reported results are credible and accurate, and that the rights, integrity, and confidentiality of trial subjects are protected’.

Investigators working with specialist and vulnerable populations should also be trained in and be aware of other guidelines. In addition, trial investigators must also be aware of their national, local and institutional regulatory requirements for the conduct of clinical trials.

An important aspect of the E6 guideline is the set of standards on the ethical and safe conduct of trials, including the assessment and reporting of adverse events (AE). Under ICH-GCP, ‘the rights, safety, and well-being of the trial subjects are the most important considerations and should prevail over interests of science and society’ (E6, 2.3).

The following glossary provides definitions on the types and seriousness of adverse events, and of other important terminology required for the understanding of the assessment and reporting of AEs.

Definitions and Terminology

Definitions of Adverse Events and Adverse Drug Reactions

Adverse Event (AE)

Any untoward medical occurrence in a patient or clinical investigation subject administered a pharmaceutical product and which does not necessarily have a causal relationship with this treatment.

An adverse event (AE) can therefore be any unfavorable and unintended sign (including an abnormal laboratory finding), symptom, or disease temporally associated with the use of a medicinal (investigational) product, whether or not related to the medicinal (investigational) product.

Serious Adverse Event (SAE) or Serious Adverse Drug Reaction (Serious ADR)

According to the ICH Expert Working Group (ICH Harmonised Tripartite Guideline 2003) a serious adverse event or reaction is any untoward medical occurrence that at any dose:

- Results in death,
- Is life-threatening (the term "life-threatening" in the definition of "serious" refers to an event/reaction in which the patient was at risk of death at the time of the event/reaction; it does not refer to an event/reaction which hypothetically might have caused death if it were more severe),
- Requires inpatient hospitalization or results in prolongation of existing hospitalization,
- Results in persistent or significant disability/incapacity,
- Is a congenital anomaly/birth defect,
- Is a medically important event or reaction.

Medical and scientific judgment should be exercised in deciding whether other situations should be considered serious, such as important medical events that might not be immediately life-threatening or result in death or hospitalization, but might jeopardize the patient or might require intervention to prevent one of the other outcomes listed in the definition above. Examples of such events are intensive treatment in an emergency room or at home for allergic bronchospasm, blood dyscrasias or convulsions that do not result in hospitalization, or development of drug dependency or drug abuse.

Adverse Drug Reaction (ADR)

In the pre-approval clinical experience with a new medicinal product or its new usages, particularly as the therapeutic dose(s) may not be established, all noxious and unintended responses to a medicinal product related to any dose should be considered adverse drug reactions.

The phrase, responses to a medicinal product, means that a causal relationship between a medicinal product and an adverse event is at least a reasonable possibility, namely, the relationship cannot be ruled out.

Regarding marketed medicinal products, the definition of an adverse drug reaction is: a response to a drug which is noxious and unintended and which occurs at doses normally used in man for prophylaxis, diagnosis, or therapy of diseases or for modification of physiological function.

Unexpected Adverse Drug Reaction and Suspected Unexpected Serious Adverse Reaction (SUSAR)

According to the ICH Expert Working Group, (ICH Harmonised Tripartite Guideline 2003):

An **Unexpected Adverse Drug Reaction** is an ADR whose nature, severity, specificity, or outcome is not consistent with the term or description used in the local/regional product labelling (e.g. Package Insert or Summary of Product

Characteristics) and therefore should be considered unexpected. When a Marketing Authorization Holder (MAH) is uncertain whether an ADR is expected or unexpected, the ADR should be treated as unexpected.

An expected ADR with a fatal outcome should be considered unexpected unless the local/regional product labelling specifically states that the ADR might be associated with a fatal outcome (ICH Harmonised Tripartite Guideline 2003).

If an unexpected adverse drug reaction is a Serious Adverse Drug Reaction, it is deemed to be a **Suspected Unexpected Serious Adverse Reaction (SUSAR)**. To be denoted as a SUSAR, according to the Central Committee on Research Involving Human Subjects (CCMO) in the Netherlands (Central Committee on Research Involving Human Subjects (CCMO) 2015), it must meet three criteria, namely:

1. The event must be serious, that is to say that irrespective of the dose the event:
    - Is fatal, and/or
    - Is life-threatening for the research subject, and/or;
    - Makes hospital admission or an extension of the admission necessary, and/or
    - Causes persistent or significant invalidity or work disability, and/or
    - Manifests itself in a congenital abnormality or malformation.
2. There must be a certain degree of probability that the event is harmful, and an undesirable **reaction to the medicinal product being** researched, regardless of the administered dosage. In other words, **there is an adverse reaction.**
3. The adverse reaction must be **unexpected.** That is to say, the nature and severity of the adverse reaction are **not in agreement with the product information as recorded in the** Summary of Product Characteristics (for an authorized medicinal product) and the Investigator's Brochure (for an unauthorized medicinal product).

Other Useful Terminology

Investigator

A person responsible for the conduct of the clinical trial at a trial site. If a trial is conducted by a team of individuals at a trial site, the investigator is the responsible leader of the team and may be called the principal investigator.

Sub-investigator

Any individual member of the clinical trial team designated and supervised by the investigator at a trial site to perform critical trial-related procedures and/or to make important trial-related decisions (e.g., associates, residents, research fellows). See also Investigator.

Sponsor

An individual, company, institution, or organization which takes responsibility for the initiation, management, and/or financing of a clinical trial.

Institutional Review Board (IRB)/Institutional Ethics Committee (IEC)

An independent body constituted of medical, scientific, and non-scientific members, whose responsibility is to ensure the protection of the rights, safety and well-being of human subjects involved in a trial by, among other things, reviewing, approving, and providing continuing review of the trial protocol and amendments and of the methods and material to be used in obtaining and documenting informed consent of the trial subjects.

Independent Data-Monitoring Committee (IDMC) (Data and Safety Monitoring Board, Monitoring Committee, Data Monitoring Committee)

An independent data-monitoring committee that may be established by the sponsor to assess at intervals the progress of a clinical trial, the safety data, and the critical efficacy endpoints, and to recommend to the sponsor whether to continue, modify, or stop a trial. The IDMC is also denoted as a data and safety management committee (DSMC) or board (DSMB).

Investigator's Brochure

A compilation of the clinical and nonclinical data on the investigational product(s) which is relevant to the study of the investigational product(s) in human subjects.

The Recording and Evaluation of Adverse Events

The sponsor is responsible for the ongoing safety evaluation of the investigational medicinal product(s). Please note, in small studies or academic-led studies, the sponsor may also be an investigator.

The sponsor must arrange systems and written standard operating procedures (SOPs) to ensure quality standards are met in the identification, documentation, grading, archiving and reporting of adverse events, and provide these SOPs to all study investigators.

Study investigators and research staff must routinely and prospectively identify any individual adverse events, record these on an adverse event case report form

(CRF), and evaluate and report the adverse event to the sponsor for evaluation. The investigator needs to evaluate the seriousness and causality between the investigational medicinal product(s) and/or concomitant therapy and the adverse event.

The sponsor is responsible for retaining detailed records of all adverse events reported by investigator(s) and performing an evaluation with respect to seriousness, causality and expectedness.

Assessment of Seriousness, Causality and Expectedness of Adverse Events

Seriousness

The reporting investigator makes the judgment as to whether the event is serious according to the definition of serious adverse event and serious adverse drug reaction (European Commission 2011).

Causality

The reporting investigator usually determines whether there is a reasonable possibility of a causal relationship. All adverse events judged by an investigator or sponsor as having a reasonable suspected causal relationship to the investigational medicinal product(s) qualify as adverse reactions.

If the investigator does not provide information on causality, the sponsor should consult the investigator for an opinion. The sponsor should not downgrade a causality assessment given by an investigator. If a sponsor disagrees with an investigator, both opinions should be provided within reports (European Commission 2011).

Expectedness

Assessment of expectedness is usually done by the sponsor according to a reference document such as (a) the investigator's brochure for a non-authorized investigational medicinal product, or (b) the summary of product characteristics for an authorized medicinal product.

Reporting of Adverse Events

Serious Adverse Events

All serious adverse events (SAEs) should be reported immediately to the sponsor except for those SAEs that the protocol or other document (e.g. Investigator's Brochure) identifies as not needing immediate reporting (ICH E6).

The immediate reports should be followed promptly by detailed, written reports. The immediate and follow-up reports should identify subjects by unique code numbers assigned to the trial subjects rather than by the subjects' names, personal identification numbers, and/or addresses. The investigator should also comply with the applicable regulatory requirement(s) related to the reporting of unexpected serious adverse drug reactions to the regulatory authority(ies) and the IRB/IEC (ICH E6).

Suspected Unexpected Serious Adverse Reactions (SUSARs)

Cases of adverse drug reactions that are both serious and unexpected are subject to expedited reporting. The reporting of serious expected reactions in an expedited manner varies among countries. Non-serious adverse reactions, whether expected or not, would normally not be subject to expedited reporting (ICH E2D).

The sponsor should expedite the reporting to all concerned investigator(s) and institutions(s), to the IRB(s) and IEC(s), where required, and to the regulatory authority(ies) of all adverse drug reactions (ADRs) that are both serious and unexpected (ICH E6).

Such expedited reports should comply with the applicable regulatory requirement(s) and with the ICH Guideline for Clinical Safety Data Management: Definitions and Standards for Expedited Reporting (ICH E6) (ICH Expert Working Group 1994).

The sponsor should submit to the regulatory authority(ies) all safety updates and periodic reports, as required by applicable regulatory requirement(s) (ICH E6).

The sponsor shall inform all involved investigators of relevant information about SUSARS that could adversely affect the safety of subjects.

Minimum Criteria for Reporting a SUSAR

SUSARs must undergo expedited reporting. At the time of initial reporting information may be incomplete. As much information as is possible should be collected for the initial report. It should at a minimum report the following: (a) a suspected investigational medicinal product, (b) an identifiable subject (e.g. study subject code number, age, sex), (c) an adverse event assessed as serious and unexpected, and for which there is a reasonable suspected causal relationship, (d) an identifiable reporter, and (e) a study protocol number where applicable.

The sponsor should report further relevant information as follow-up reports, and in certain cases, it may be appropriate to conduct follow-up of the long-term outcome of a particular reaction. Follow-up reports must provide all the appropriate information required for an adequate analysis of causality.

In blinded (masked) trials, as a general rule treatment codes should be broken for that specific subject by the sponsor before reporting a SUSAR to competent authorities and ethics committees.

Types of information reported to determine causality can include the description of the reaction, criteria for assessing seriousness, signs and symptoms, specific diagnosis of the reaction, onset date and time of the reaction, stop data and time or duration of the reaction, dechallenge and rechallenge information, diagnostic tests and laboratory data, setting, outcome including recovery and sequelae relatedness of product to reactions and events (ICH-E2D). For a fatal outcome a cause of death, relevant autopsy or post-mortem findings, should also be reported.

Reporting Non-serious Adverse Events and/or Laboratory Abnormalities

Adverse events and/or laboratory abnormalities identified in the protocol as critical to safety evaluations have to be reported to the sponsor according to reporting requirements and timeframes in the study protocol. The sponsor must keep detailed records of all adverse events reported to him/her by the investigator(s) (European Commission 2011).

Conclusion

In conclusion, all researchers must ensure that the rights, safety, and well-being of the trial subjects are the most important considerations, and these prevail over interests of science and society. Guidelines such as ICH-GCP amongst others, along with regulatory bodies and committees, provide the standards and framework for this to occur, and for the safety of subjects to be monitored by the sponsor and independently of the sponsor by IDMCs, IRBs, IECs and Governmental agencies. By centralizing and standardizing the reporting of events to the sponsor, this allows the sponsor and others to have an overall perspective of the rate and types of events across sites. It allows the sponsor to monitor sites with high adverse event rates more closely, to make any protocol amendments to improve safety, and to collate and publish adverse events so that the safety profile of the medication within subpopulations can be reported in addition to the effectiveness of the test drug.

References

Central Committee On Research Involving Human Subjects (CCMO) (2015) SAEs/SUSARs [Online]. http://www.ccmo.nl/en/saes-susars

European Commission (2011) Communication from the commission – detailed guidance on the collection, verification and presentation of adverse event/reaction reports arising from clinical trials on medicinal products for human use (CT-3) – 2011_c172_01_en.pdf [Online]. http://ec.europa.eu/health/files/eudralex/vol-10/2011_c172_01/2011_c172_01_en.pdf

ICH Expert Working Group (1994) *ICH TOPIC E2A – E2A*_Guideline.pdf [Online]. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2A/Step4/E2A_Guideline.pdf

ICH Harmonised Tripartite Guideline (2003) *ICH E10 – E2D_Guideline.pdf* [Online]. http://www.ich.org/fileadmin/Public_Web_Site/ICH_Products/Guidelines/Efficacy/E2D/Step4/E2D_Guideline.pdf

International Conference on Harmonisation (2015) ICH official web site: ICH [Online]. http://www.ich.org/home.html

Chapter 11
Research Ethics and N-of-1 Trials

Andrew Crowden, Gordon Guyatt, Nikola Stepanov, and Sunita Vohra

Abstract Some N-of-1 trials are conducted as part of clinical care, others are developed as research. For those that are research, unless they are deemed exempt from formal review, a relevant Human Research Ethics Committee or Institutional Review Board should review specific projects before they are approved. N-of-1 trials should also be authorized by institutions before commencing. The level of risk to the patient/participant should guide and determine whether a particular project is exempt from review, subject to a low/negligible risk review, or should be reviewed by a full committee. Research ethics reviewers must develop a heightened ethical sensitivity toward ensuring that a misguided approach to N-of-1 review does not occur. Clinical researchers, institutions and research review committees, should recognize the continuum of clinical care and clinical research, in order to set and act from explicit standards which are consistent with the clinical practice – clinical research continuum.

Keywords Research ethics • N-of-1 trials • Clinical research continuum • Bias • Respect • Merit • Integrity • Justice • Beneficence • Informed consent • Risk assessment • Intent • Institutional Review Boards • Human Research Ethics Committees

A. Crowden (✉)
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: a.crowden@uq.edu.au

G. Guyatt
McMaster University, Hamilton, Canada
e-mail: guyatt@mcmaster.ca

N. Stepanov
School of Medicine, The University of Queensland, Brisbane, Australia
e-mail: n.stepanov@uq.edu.au

S. Vohra
Department of Pediatrics, University of Alberta, Edmonton, Canada
e-mail: svohra@ualberta.ca

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_11

Introduction

The same ethical values that underpin clinical care also underpin clinical research: respect for human beings, research (or clinical) merit and integrity, justice, and beneficence. The specific values that are the foundation for ethical relationships between researchers and research participants are intrinsically connected to widely acknowledged values for effective ethical therapeutic relationships between clinicians and patients. The weight placed on particular values in clinical care and research may, however, differ depending on the context. Researchers and research reviewers, for instance, tend to place a higher value on respect for persons and autonomy, and therefore focus on the consent process and associated documentation (Appelbaum et al. 1987; Faden et al. 1986; Stepanov and Smith 2013: Stepanov 2014). This necessitates ensuring that potential participants are provided with information and consent forms that clearly articulate the justifications for the proposed research, including its scientific merit and integrity (Kerridge et al. 2013). While the standards applied to the delivery of clinical care and the conduct of research may appear to be growing more divergent over time, in practice the boundary between clinical care and clinical research often cannot be clearly differentiated (Lewens 2006; Kottow 2009).

Perhaps there is no better illustration of the close connection between clinical practice and clinical research than that evidenced by recent developments in innovative N-of-1 trials (Guyatt 1996). Analysis of the ethical dimensions of these trials is pertinent and instructive. We will begin this brief exploration of the ethics of N-of-1 trials and the nature of the relationship between clinical care and clinical research by highlighting certain aspects from the N-of-1 trials story.

Reducing Bias in Clinical Research and the Development N-of-1 Trials

To ensure accurate results, one needs methods of inquiry that reduce potential bias (Keech et al. 2007). Medical history is littered with misleading results. Many once popular, but now discarded treatments, were previously thought to be effective, but are now known to be useless or harmful (Brignardello-Petersen et al. 2014). Accordingly, the last 50 years has seen the development of increasingly sophisticated strategies for minimizing bias in establishing intervention effectiveness and thus avoiding misleading results.

Even the most rigorously conducted randomized trials are limited by variability in patient responses. Positive trials do not mean that every patient benefits. Clinicians recognize this variability in response and conduct traditional trials of therapy in which they offer an intervention to patients and then monitor patients' response. Such conventional trials of therapy may, however, result in misleading conclusions because of natural history (the patient may be destined to improve irrespective of therapy); placebo effects; patients and physicians' expectations; and the patients' desire to please.

Applying the same safeguards by applying, to trials of therapy in individual patients, strategies such as randomization and blinding common in large Randomized Control Trials (RCTs) can reduce bias that results in misleading inferences. Psychologist researchers have long used such methods in what they have called 'single case' or 'single subject research' (Tate et al 2013). When applied to real-life treatment decisions, clinical epidemiologists called the strategy 'N-of-1 Randomized Trials', N-of-1 denoting that each randomized trial involved only a single patient. Today we tend to call them N-of-1 trials, and they may occur in one patient, a series of patients, with varying designs (with or without blinding or randomization), and as part of clinical care or research.

In general terms, the most commonly used N-of-1 design is based on multiple pairs of active/placebo, high/low dose, or first drug/alternative drug combinations, the order of each pair determined by random allocation. When N-of-1 trials are undertaken as part of clinical care, the clinician and patient enter a partnership. The clinician and patient monitor treatment targets (usually directed at specific patient complaints or symptoms) in a blinded fashion, on a regular predetermined schedule (Guyatt et al. 1986; Gabriel and Normand 2012). N-of-1 trials pose an ideal setting for patient-centered care, as patients can help determine the outcomes of greatest interest to them, and participate in recording their status on these outcome measures. The trial continues as long as the patient and clinician agree that they need more information to get a definitive answer regarding the efficacy, superiority or side effects of the treatment, or until the patient or clinician decides, for any other reason, to end the trial.

Ethics and N-of-1 Randomized Trials

The ethical implications of N-of-1 trials became evident when early developers of N-of-1 trials considered the circumstances in which to conduct such a trial. A typical experience is outlined in the following case:

Susan, an experienced clinician, faces Derek, a patient suffering from chronic obstructive pulmonary disease who remains with troubling dyspnea despite treatment with inhaled tiotropium, inhaled steroids, and as-needed inhaled salbutamol. Susan considers adding an oral theophylline, but notes that it is often associated with adverse effects, and that though randomized trials have shown that on average it reduces dyspnea in such patients, there is considerable variability in patient response. Susan decides to prescribe theophylline (evidence shows will help some, but not all the patients, to whom it is offered). Derek, the patient before her may be one of those who benefit, or one who receives no symptom relief but only treatment side-effects. *How might Susan handle the situation?*

One option would be to prescribe theophylline and leave it at that. For treatments such as theophylline in COPD where RCTs have shown overall benefit, clinicians adopting this approach can at least be confident that on balance, their patients are more likely to benefit than not. However, when trial results suggest large variability in response to treatment, or when patients differ substantially from those enrolled in the available randomized trials, the benefit for the individual patient will be uncertain. Thus, rather than simply prescribing therapy, the clinician may choose to conduct a conventional trial of therapy. If that trial shows apparent benefit, the intervention (typically a drug) is continued; if not, it is stopped.

Physicians who are aware of the aforementioned sources of bias may not, however, be satisfied with conventional trials of therapy. N-of-1 trials potentially minimize the bias of a conventional trial of therapy, allowing much greater confidence in inferences regarding individual benefit and may result in changes in treatment decisions up to 35 % of the time (Guyatt et al 1986). Thus, because the conclusion for treatment choice will be far less likely to be spurious, clinicians such as Susan in the scenario above, may decide it is in their patient's best interest to conduct an N-of-1 trial.

Having made this judgment, how should the clinician proceed? Susan could explain the uncertainties regarding the treatment decision to Derek. If Derek understands the concept of an N-of-1 trial, is competent to understand potential risks and benefits, and is willing to be involved, Susan and Derek, in partnership, can plan and conduct an N-of-1 trial.

Or can they? Should Susan apply to a Human Research Ethics Committee or Institutional Review Board for approval? Should the institution authorize the research? Is authorization required from any other body such as a government health department (often these are location specific, for example the *Therapeutic Goods Administration* in Australia, or in Canada, *Health Canada*)?

Human Research Ethics Committees, Review Boards and N-of-1 Trials

When investigators at McMaster University proposed the first N-of-1 trials in the early 1980s, their Institutional Review Boards (IRB) didn't see the need for research ethics review. The McMaster IRB viewed N-of-1 trials as 'optimal clinical care' and not research.

Subsequent experience has been varied. Some IRBs or Human Research Ethics Committees consider N-of-1 trials to be research, while others conclude that they can be either research or clinical care. If the primary goal is to test treatments for the purpose of contributing to further knowledge about how to manage and treat a condition in average patients, then the N-of-1 trial may be most appropriately considered research. However, if the primary goal is to improve the care of the individual patient, then the N-of-1 trial may be appropriately considered clinical care (Punja et al. 2014). What is the correct view? At what point do clinical activities become research? Are N-of-1 trials research?

N-of-1 trials illustrate the difference between the standards we set for clinical care and the standards we set for research. The difference would not be a problem if there were a clear boundary between clinical and research activities. Systematic clinical observation, continuous quality improvement and clinical research are overlapping activities. These activities are on a continuum and the ethical standards we apply should reflect that continuum.

As an illustration of the continuum between quality improvement efforts and clinical research consider the following. A clinician wishes to monitor the extent to which she is successful in achieving full vaccination for all children in her practice. No one is likely to suggest that she is conducting research, or that she had better appear before an institutional review board or risk subsequent censure from her colleagues when her clandestine research activities are brought to light. What if she wishes to conduct her monitoring in collaboration with a number of colleagues with one goal: to ultimately compare how well each of them is doing? What if, as a group, these physicians negotiate with the local public health department for a public health nurse to help them establish registries of their patients with systematic reminders to help achieve full vaccination? What if they now monitor, in a before-after fashion, the extent to which the intervention of the health department improved the vaccination rate? Finally, what if the group decides to publish the results of their experience, in the hopes that they might be beneficial to others? At what point in this series of possible activities related to vaccination does the transition from quality assurance not requiring research ethics oversight to a research activity requiring such oversight occur?

As we have previously noted, the specific values that are the foundation for ethical relationships between researchers and research participants are intrinsically connected to widely acknowledged foundational values for effective ethical therapeutic relationships between clinicians and patients. These values include respect for human beings, research (or clinical) merit and integrity, and justice and beneficence (Beauchamp and Childress 2001). These values tend to be applied at a higher standard in research, and the ranking of particular values may differ depending on the context. Whether one considers Susan's proposed N-of-1 trial the delivery of optimal clinical care (requirement only that the patient understands and consents) versus research (the necessity for review by a human research ethics committee or IRB), illustrates the nature of the problem.

Most of us recognize that we must treat all rational beings, or persons, never merely as a means, but always as ends. There are strong reasons to accept some version of Immanuel Kant's famous formula for humanity, the so-called 'consent principle'. Sometimes clinicians may find that consent seems too demanding. However in most cases, clinicians agree that it is wrong to act in ways to which a person might not rationally consent (Parfitt 2013).

In relation to a proposed clinical practice intervention, or a choice about participation in a research project, it is generally accepted that the patient or potential participant should make a decision. Relevant information is disclosed by the clinician/researcher. If patient participants are competent to understand the information, they can voluntarily choose to make an informed decision about whether to accept treatment or participate in the research project. Good in theory, but what about practice? Let's return to Susan, our clinician, Derek her patient, and their N-of-1 trial partnership.

If Susan had decided to undertake a conventional trial of therapy believing it would benefit Derek, she would obtain consent in her usual way and begin. For most clinicians, this would be implied consent only. In other words, the clinician would suggest the intervention and, in the absence of objections from the patient, write the prescription and presume that the patient will proceed accordingly. However, for good reasons, she has chosen to conduct a trial of therapy in partnership with Derek in a much more rigorous manner, with the idea of reaching a more trustworthy assessment of benefit. In one view, apparently, by being more rigorous and explicitly involving the patient in the decision, the clinician, rather than conscientiously discharging her ethical responsibilities, is taking on additional ones.

Clinicians considering N-of-1 trials have to think carefully about their intent. If they are trying to improve the individual patient's care, then the N-of-1 trial can be conducted as part of clinical practice. If their intent is to develop knowledge to benefit others (i.e. their primary intent is research), this may influence what patients are offered or how their outcomes are measured. Under these circumstances, in order to obtain research approval and institutional authorization before the therapy can begin they must write a research protocol and complete an ethics application as well as meeting any other local site authorization requirements. A research ethics review board will consider the proposal, the consent procedure, and scrutinize other relevant underlying values.

As John Lantos has rightly claimed, a researcher is evaluating therapy, while a clinician who conducts a conventional trial of therapy makes her inferences, because of the biases we have previously noted, based on very imperfect information. It does seem odd that Susan (a responsible clinician who understands the principle of evidence-based practice) requires external review and regulation because she chose to be more responsible in ascertaining what is best for her patient than colleagues who would conduct conventional far less rigorous trials of therapy (Lantos 1994). This is the double standard on consent to treatment/research that may frustrate clinicians and potentially disadvantage patients. The clinician who is prepared to admit uncertainty about therapy, and address a need to for safeguards against bias, is subject to more stringent rules than clinicians who don't.

We regard this double standard as illogical and indefensible. The onus must be on us all to ensure that participation in N-of-1 trials, or indeed any clinical care procedure, is not always presented as a high-risk endeavor (Evans et al. 2013, p. 166). The double standard anomaly is most obvious in the consent example. The double standard anomaly is also relevant to other key values. For example, merit and integrity, justice, beneficence, like respect for persons, are all reviewed and regulated in a more robust manner in a research context. Clinical researchers are aware of this anomaly. Research approving and authorizing bodies should be too.

This brings us to the last key issue with N-of-1 trials. How should they be reviewed? Should N-of-1 trials be exempt from research ethics review, or not? The answer to this question is as expected – it depends!

Reflecting on the differences between clinical care, quality improvement and human research is a useful exercise prior to undertaking an N-of-1 trial. Some N-of-1 trials may not need research ethics review, while others do. For instance, an N-of-1 trial done in the clinical context may not be research any more than the conventional trial of therapy. If the intent is identical, the only difference may be that the former is more rigorous and likely to lead to accurate inferences and the latter is less rigorous and more likely to lead to spurious inferences.

Deciding whether an N-of-1 trial requires a formal research ethics review is dependent primarily on potential risks and benefits, and how those risks and benefits are conveyed to prospective participants. We know that the research ethics review process, including by relevant bodies like ethics committees, considers factors such as the robust nature of the evidence used to predict actual and potential participant risks, the quality of the literature review; and the reported outcomes of any studies that informed the design of the N-of-1 study being reviewed. Novice N-of-1 clinician researchers will benefit from consulting a relevant N-of-1 user guide. For instance the Agency for Healthcare Research and Quality's *Design and Implementation of N-of-1 Trials: A User's Guide* provides useful checklists (Kravitz and Duan 2014).

It is not surprising that it has become more common internationally for the specific level of risk to participants to determine the type and comprehensiveness of research ethics review that is required. In Australia, national guidelines and 'best practice' have determined that it is the assessment of risk, the likelihood for harm, discomfort or inconvenience that should determine the type of review that is required for a particular project. In this regard N-of-1 trials are treated the same as any other research project. If risk is low, they may be exempt from review. N-of-1 trials that are exempt would include those projects that are conducted strictly to optimize treatment for the individual. Such trials may require institutional site assessment and authorization, but not research ethics approval. Where risk assessment indicates that the research has low risk or even negligible risk, an N-of-1 project is not usually reviewed by a full research ethics committee, though some form of expedited review is often employed.

Different nations may have different requirements, and researchers and clinicians have legal and ethical obligations to make themselves familiar with the requirements in their jurisdiction, and to comply with those standards. Some N-of-1 trials will be higher than minimal risk, or more than the risk related to everyday life. Still the key issue is whether they are at higher risk than the real alternative, which is the conventional trial of therapy.

When compared to N-of-1 trials, the risks to the patient of conventional trials of therapy or just handing out treatment without monitoring will always be greater. We cannot think, in this comparison, of an N-of-1 trial situation where beyond minor discomfort there may be a real increased risk to participants of physical, psychological, social, economic or even legal risk.

There may however be risks to non-participants including distress to family members that should be considered. For example focuses on ethical sensitivity and considered care are important, especially when indigenous people are involved (Crowden 2013). However again, it is likely that conventional therapy will have greater risk than N-of-1 trials.

Conclusion

Some N-of-1 trials are conducted as part of clinical care, others are developed as research. For those that are research, unless they are deemed exempt from formal review, a relevant Human Research Ethics Committee or Institutional Review Board should review specific projects before they are approved. N-of-1 trials should also be authorized by institutions before commencing. The level of risk to the patient/participant should guide and determine whether a particular project is exempt from review, subject to a low negligible risk review, or should be reviewed by a full committee.

Research ethics reviewers must develop a heightened ethical sensitivity toward ensuring that a misguided approach to N-of-1 review does not occur. They must ensure that the identified double standards between clinical care and clinical research do not persist. Clinical researchers, institutions and research review committees, should recognize the continuum of clinical care and clinical research, in order to set and act from explicit standards which are consistent with the clinical practice – clinical research continuum. We should recognize that N-of-1 trials are a better ethical alternative when compared to conventional therapy. The notion that optimal clinical practice in the form of an N-of-1 clinical trial requires greater oversight than usual suboptimal clinical practice is indefensible.

References

Appelbaum PS, Lidz C, Meisel A (1987) Informed consent. Oxford University Press, New York

Beauchamp T, Childress J (2001) Principles of biomedical ethics. Oxford University Press, New York

Brignardello-Petersen R, Ioannidis JPA, Tomlinson G, Guyatt G (2014) Surprising results of randomized trials. In: Guyatt G, Meade MO, Cook DJ, Rennie D (eds) Users' guides to the medical literature: a manual for evidence-based clinical practice, 2nd edn. McGraw-Hill, New York

Crowden A (2013) Chapter 6: ethics and indigenous health care: cultural competencies, protocols and integrity. In: Hampton R, Toombs M (eds) Indigenous Australians and health: the wombat in the room. Oxford University Press, South Melbourne, Victoria, Australia, pp 114–129

Evans I, Thornton H, Chalmers I, Glasziou P (2013) Testing treatments: better research for better healthcare, 2nd edn. Pinter &Martin Ltd., London

Faden R, Beauchamp T, King N (1986) A history and theory of informed consent. Oxford University Press, New York

Gabriel SE, Normand SL (2012) Getting the methods right – the foundation of patient-centered outcomes research. N Engl J Med 367(9):787–790

Guyatt G (1996) Clinical care and clinical research: a false dichotomy. In: Daly J (ed) Ethical intersections: health research, methods and researcher responsibility. Allen and Unwin, Sydney, pp 66–73

Guyatt GH, Sackett DL, Taylor DW et al (1986) Determining optimal therapy-randomised trials in individual patients. N Engl J Med 314:889–892

Keech A, Gebski V, Pike R (2007) Interpreting and reporting clinical trials. A guide to the consort statement and the principles of randomised controlled trials. Australian Medical Publishing, Sydney

Kerridge I, Lowe M, Stewart C (2013) Ethics and law for the health professions. The Federation Press, Leichardt

Kottow M (2009) Clinical and research ethics as moral strangers. Arch Immunol Ther Exp (Warsz) 57:157–164

Kravitz R, Duan N (eds) (2014) Design and Implementation of N-of-1 Trials: a user's guide, Agency for healthcare research and quality, US Department of Health and Human Services, Feb 2014

Lantos J (1994) Ethical issues – how can we distinguish clinical research from innovative therapy? Am J Pediatr Hematol/Oncol 16:72–75

Lewens T (2006) Distinguishing treatment from research: a functional approach. J Med Ethics 32:424–429

Parfitt D (2013) On what matters, vol 1. Oxford University Press, Oxford/New York, p 178

Punja S, Vohra S, Eslick I, Duan N (2014) An ethical framework for N-of-1 trials: clinical care, quality improvement, or human subjects research? In: Kravitz RL, Duan N (eds) Design and implementation of N-of-1 trials: a user's guide. AHRQ Publication No. 13(14)-EHC122-EF. Rockville, pp 13–22. www.effectivehealthcare.ahrq.gov/N-1-Trials.cfm

Stepanov N (2014) Questioning the boundaries of parental rights: exploring children's rights, the best interests standard, and parental consent to paediatric non-therapeutic experimental research. Doctor of Philosophy PhD, The University of Melbourne

Stepanov N, Smith MK (2013) Double standards in special medical research: questioning the discrepancy between requirements for medical research involving incompetent adults and medical research involving children. J Law Med 21:47–52

Tate RL, Perdices M, Rosenkoetter U et al (2013) Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item risk of bias in N-of-1 trials (RoBiNT) scale. Neuropsychol Rehabil 23(5):619–638

Chapter 12
Statistical Analysis of N-of-1 Trials

Kerrie Mengersen, James M. McGree, and Christopher H. Schmid

Abstract This chapter discusses some techniques for exploratory data analysis and statistical modelling of data from N-of-1 trials, and provides illustrations of how statistical models and corresponding analyses can be developed for the more common designs encountered in N-of-1 trials. Models and corresponding analyses for other designs, perhaps involving different nesting of treatments, order and blocks, can be developed in a similar manner. The focus of this chapter is on continuous response outcomes, that is, numerical response data. The chapter is presented in tutorial style, with concomitant R code and output provided to complement the description of the models. Mixed effects models are also discussed. Such models can be extended to account for a variety of factors whose effects can be considered as random draws from a population of effects. A taxonomy of relevant statistical methods is also presented. This chapter is aimed at readers with some background in statistics who are considering an analysis of data from an N-of-1 trial in the R package.

Keywords Correlated measurements • Exploratory data analysis • Goodness-of-fit • Linear models • N-of-1 trials • Nonparametric methods • Statistical modelling • The R-package • Treatment effects

Introduction

Once data from an N-of-1 trial have been collected, they need to be analyzed. The methods and models adopted for analysis will depend on the way in which the trial has been designed and the aim of the analysis.

K. Mengersen (✉) • J.M. McGree
Mathematical Sciences, Queensland University of Technology, Brisbane, Australia
e-mail: k.mengersen@qut.edu.au; james.mcgree@qut.edu.au

C.H. Schmid
Department of Biostatistics and Center for Evidence Based Medicine,
Brown University, Providence, RI, USA
e-mail: christopher_schmid@brown.edu

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_12

The purpose of this chapter is to review a range of statistical approaches for a typical, single N-of-1 trial. In order to enhance the practical applicability of the methods presented here, we focus on a simulated study that is similar to that used by Schmid and Duan (2014), and provide details of the analysis in the statistical software package R. For further reading, the reader is directed to Chap. 16, which deals with the aggregated analysis of many N-of-1 trials.

Simulated Case Study

Suppose that the trial aims to evaluate an outcome associated with two treatments. The patient is exposed to each treatment in six blocks (replicates). An example dataset for this study is given below (Table 12.1). It involves six blocks of two time periods each during which the patient receives each treatment in randomized order. We note that the statistical techniques implemented in this chapter are not specific to the particular design of this N-of-1 trial, meaning they could be applied to analyze data from a range of different experimental designs.

There are a few ways to set up the data in the software package R. Two of these are as follows.

- Open R and specify a working directory that identifies the location for the data and the results. For example, in Windows, if the location is a folder called ‘example’ within the folder ‘trials’ on the C drive of the computer, this would be achieved with the command:

  ```
  setwd("C:/trials/example")
  ```

Table 12.1 Simulated case study example dataset

Time period	Block	Treatment	Order	Outcome
1	1	1	1	31
2	1	2	2	35
3	2	1	2	28
4	2	2	1	39
5	3	1	1	32
6	3	2	2	39
7	4	1	2	36
8	4	2	1	37
9	5	1	1	38
10	5	2	2	41
11	6	1	2	39
12	6	2	1	39

- *Option 1:*
  Type the data in an Excel spreadsheet and save it as a csv file, say “data.csv”, in the ‘example’ directory. Then read the file into R as an object called ‘egdat’, say, using the command:

  ```
  egdat = read.csv(”data.csv”)
  egdat = data.frame(egdat)
  attach(egdat)
  ```

- *Option 2:*
  If the dataset is sufficiently small, type it directly into R:

  ```
  time = seq(1,12)
  block = c(1,1,2,2,3,3,4,4,5,5,6,6)
  treat = rep(c(1,2),6)
  order = rep(c(1,2,2,1),3)
  outcome= c(31,35,28,39,32,39,36,37,38,41,39,39)
  egdat = data.frame(time, block, treat, order, outcome)
  attach(egdat)
  ```

  Attaching the data frame allows you to reference the variables inside it without having to also reference the data frame. Now check the number of rows and columns of egdat and what data are in it:

  ```
  dim(egdat)
  egdat
  head(egdat)
  ```

  By default, any numbers that appear within the dataset will be considered within R as numeric. For variables that are actually factors, such as ‘block’, it is important to set these as factor variables. This can be achieved for the appropriate variables as follows:

  ```
  block = as.factor(block)
  ```

  The output of the command ‘head (egdat)’ shows data from the first six observations would be

  ```
       time block treat order outcome
  1     1      1     1     1      31
  2     2      1     2     2      35
  3     3      2     1     2      28
  4     4      2     2     1      39
  5     5      3     1     1      32
  6     6      3     2     2      39
  ```

Taxonomy of Statistical Methods

Schmid and Duan (2014) provides details of a range of statistical methods used for analysis of N-of-1 trials. These methods are represented as a decision tree in Fig. 12.1. This chapter elaborates on some of these approaches. Note that other related methods, such as the treatment alone model using a *t*-test, or an ordered categorical model for scaled outcomes, may also be appropriate, depending on the design, data and the intended inference.

Exploratory Analysis

It is useful to conduct a preliminary evaluation of the data, using plots and summary statistics. Plots of the data can be obtained using the following R commands.

```
par(mfrow=c(1,2))
plot(as.numeric(block[treat==1]),outcome[treat==1],
type="l", ylim=c(25,45),xlab="Block",ylab="Outcome",main="(a)")
lines(block[treat==2],outcome[treat==2], lty=2)
legend(3,45,c("'-'  Treat=1","'--' Treat=2"))
plot(outcome[treat==1], outcome[treat==2],
xlim=c(28,41), ylim=c(28,41),xlab="Outcome for treatment
1",ylab="Outcome for treatment 2",main="(b)")
lines(c(30,41),c(30,41))
```

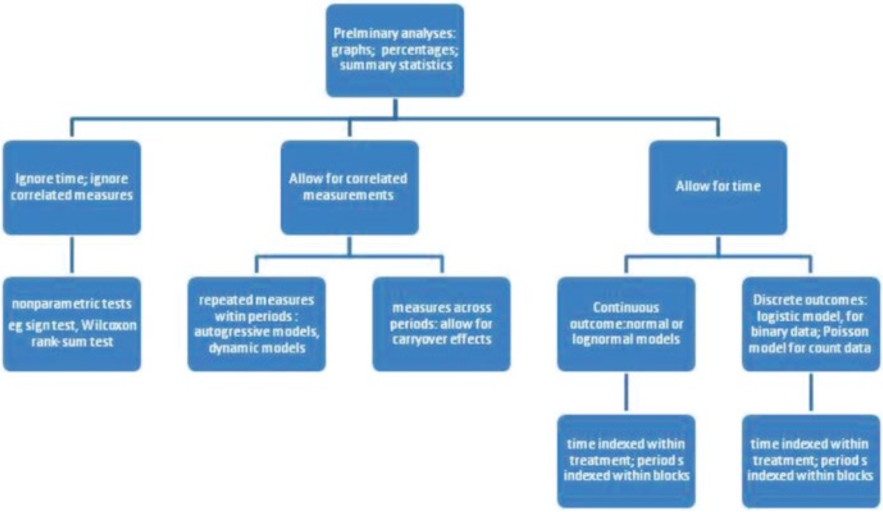

Prelminary analyses: graphs; percentages; summary statistics
Ignore time; ignore correlated measures
Allow for correlated measurements
Allow for time
nonparametric tests eg sign test, Wilcoxon rank-sum test
repeated measures witin periods: autogressive models, dynamic models
measures across periods: allow for carryover effects
Continuous outcome:normal or lognormal models
Discrete outcomes: logistic model, for binary data; Poisson model for count data
time indexed within treatment; periods indexed within blocks
time indexed within treatment; periods indexed within blocks

Fig. 12.1 Decision tree of statistical methods for analysis of N-of-1 trials

The resultant plots are shown in Fig. 12.2 and are described as follows. The left hand panel shows the outcome values for the two treatments (1: solid line; 2: dotted line), for each block (indicated on the horizontal axis). It appears that treatment 2 is better than treatment 1, although the magnitude of this improvement is not clear and appears to depend on the block and/or time.

The right hand panel shows the six pairs of outcome values plotted by treatment (1: horizontal axis; 2: vertical axis). The solid line indicates where the outcome for treatment 1 would be equal to the outcome for treatment 2. All of the points are above this line, indicating that treatment 2 is better than treatment 1.

It is important to note that the plot does not take into account other variables such as order. Similar plots could be drawn to visually evaluate the association and effect of order on the outcome.

It is also interesting to plot a histogram and an empirical density of the outcome, see Fig. 12.3. Although not shown here, one could 'color' (or otherwise identify) the histogram bars to indicate the outcomes associated with different time periods, blocks, treatments and order. This can facilitate a general evaluation of the contribution of these factors; for example if all of the outcomes for treatment A are at one end of the plot, then treatment B would appear to be better than treatment A. This is indeed shown in Fig. 12.3b, where the density for all of the outcomes (solid line) is contrasted with the density for treatment 1 (wide dotted line) and treatment 2 (taller dotted line, showing that these values are more concentrated and generally larger than the values for treatment 1).

Note that, as above, these plots are based on very small numbers so it is important not to read too much into them. They are visual inspections only, and not formal statistical tests.

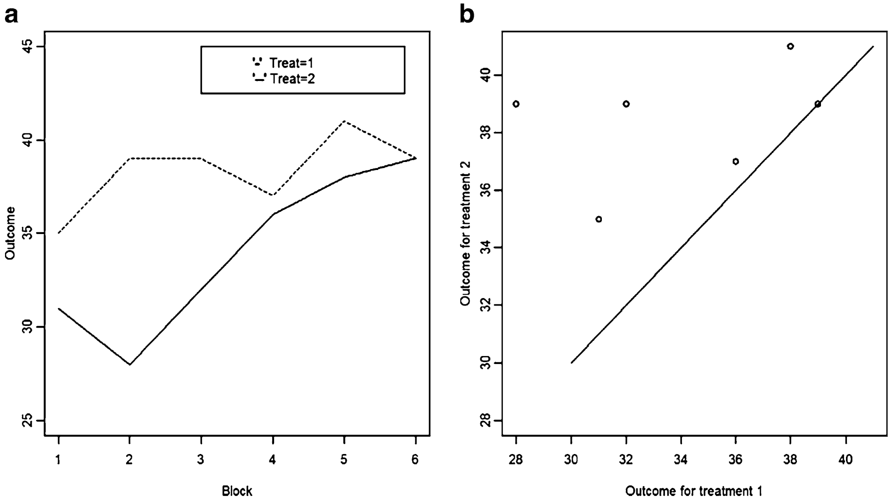
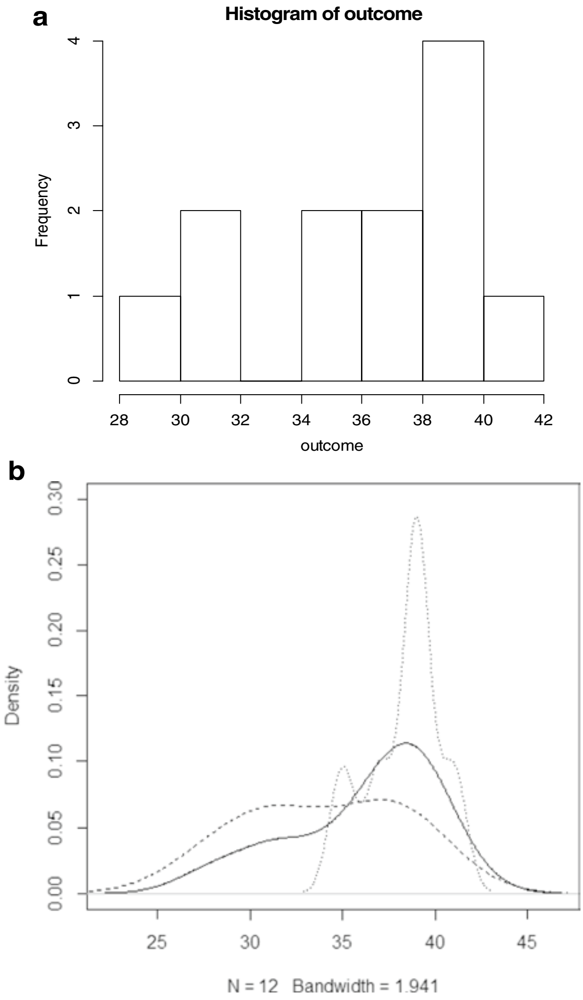


Fig. 12.2 Assessing preliminary outcomes using plots of (**a**) the outcomes of treatment 1 and treatment 2, and (**b**) direct comparison of the outcomes of treatment 1 and treatment 2 against a line indicating identical treatment effect

Fig. 12.3 Using a histogram (**a**) and empirical density of different treatment outcomes (**b**) to estimate treatment effects



```
hist(outcome)
plot(density(outcome),main="",ylim=c(0,0.30))
lines(density(outcome[treat==1]), lty=2)
lines(density(outcome[treat==2]), lty=3)
```

Overall summary statistics for the outcome can also be obtained as part of the exploratory analysis.

```
summary(outcome)
summary(outcome[treat==1])
summary(outcome[treat==2])
```

These commands display the minimum, 1st quartile, median, mean, 3rd quartile and maximum:

```
Outcome:
 Min      1st Qu    Median    Mean    3rd Qu    Max.
 28.00    34.25     37.50     36.17   39.00     41.00
Outcome for Treatment=1:
 Min.     1st Qu    Median    Mean    3rd Qu    Max.
 28.00    31.25     34.00     34.00   37.50     39.00
Outcome for Treatment=2:
 Min.     1st Qu    Median    Mean    3rd Qu    Max.
 35.00    37.50     39.00     38.33   39.00     41.00
```

The differences between outcome values for treatments 1 and 2 within each block can also be calculated. Since it is a small dataset, and for exposition, we do this manually:

```
diff=c(35-31, 39-28, 39-32, 37-36, 41-38, 39-39)
diff [1]  4 11  7  1  3  0
```

All of the values are positive, indicating that treatment 2 appears to be better than treatment 1, ignoring time and order.

Nonparametric Methods

The sign test and the Wilcoxon signed rank test (or the Mann–Whitney test) are two common nonparametric tests that can be used to test for differences between the median outcomes for the two different treatment groups. Note that although the Wilcoxon test is more informative than the sign test because it uses the ranks of the outcome values in addition to their sign, i.e. whether they are above or below the median), both tests ignore the other variables (such as time and order).

The R commands for the sign test below are in a library called BSDA. This package needs to be installed (e.g. using the 'Packages', 'Install Packages' menus in R) and attached (e.g., using the 'Packages', 'Load Packages' menus or the command 'library(BSDA)'. See R user information for more details.

```
SIGN.test(diff)
wilcox.test(outcome[treat==1], outcome[treat==2])
```

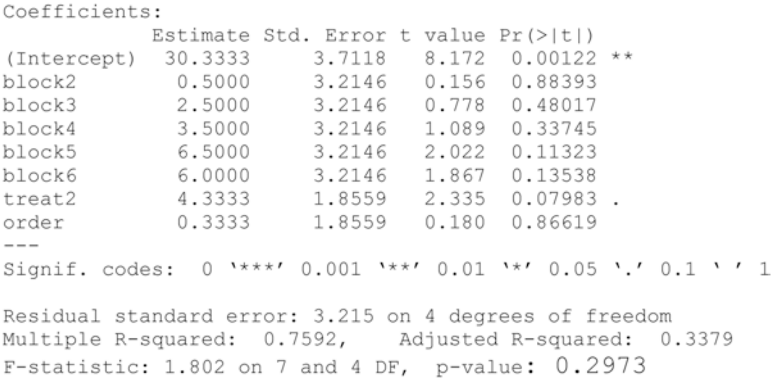
These commands test whether the median difference is equal to zero. The sign test returns a p-value of 0.06 and the Wilcoxon test returns a p-value of 0.07. Note that with such a small dataset, these values, and the test itself, may have low statistical power to detect reasonably sized differences between treatment groups. As each p-value is between 0.05 and 0.1, there is some evidence to indicate a true difference between the two treatments, but as above this still ignores the possible effect of time and order.

Linear Models

The simplest linear model describes the variation in the outcome, *y,* as a function of the variation in blocks, treatments and order. This can be easily expressed in R as follows. Here, 'outlm' is the object that holds the results of the linear model analysis. The command 'summary' then provides a summary of these results.

```
outlm = lm(outcome ~ block + treat + order)
summary(outlm)
```

This provides the following output.



```
Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  30.3333     3.7118   8.172  0.00122 **
block2        0.5000     3.2146   0.156  0.88393
block3        2.5000     3.2146   0.778  0.48017
block4        3.5000     3.2146   1.089  0.33745
block5        6.5000     3.2146   2.022  0.11323
block6        6.0000     3.2146   1.867  0.13538
treat2        4.3333     1.8559   2.335  0.07983 .
order         0.3333     1.8559   0.180  0.86619
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.215 on 4 degrees of freedom
Multiple R-squared:  0.7592,    Adjusted R-squared:  0.3379
F-statistic: 1.802 on 7 and 4 DF,  p-value: 0.2973
```

This output tells us that there is no significant effect (on the outcome values) due to differences in blocks and due to order, but that there is some evidence of a significant effect due to treatment. Multiple R-squared shows that about 76 % of the variability in the data is explained by the model. However, a much smaller adjusted R-squared value is seen (34 %). This suggests that the model is overfit. Indeed, there are 12 observations and 8 terms in the model. Generally, it is preferable to describe as much variability in the data as possible but with as fewer model terms as possible. Therefore, these results are not satisfactory, and further exploration into the development of a parsimonious statistical model would typically follow. However, it is important to understand that these nonparametric tests and the above linear statistical model do not take into account the nested structure of the data nor do they account for other potentially important features of the study, such as possible linear time trend (rather than by block) that may be present in the data, and/or autocorrelation of errors that is typically seen in data that are collected over time. These are fundamental flaws that render these approaches inappropriate for inference. In the next sections, we consider statistical techniques that are appropriate to analyze these data.

Taking Account of the Trial Design

The above discussion illustrates the way in which the statistical model and corresponding analysis are developed to reflect the design of the trial, in particular, the way in which order or treatment is nested within blocks. For example, it may be useful to consider outcomes for a particular treatment within a specified block. This can be analyzed using the same model without order, that is, by using the following R command

```
outlm <-lm(outcome~block + treat)
```

Alternatively, a model that describes the nested structures in the dataset, and the corresponding analysis of variance, can be written in R as follows. Note that the slash in the Error term indicates the nested structure of the design. Also, the use of the term ‘nesting’ of treatment within block differs from the usual use of nesting, where the nested factor levels only make sense within the higher order factor. Here, order is nested within block unless one defines order as always the first or second treatment in a block and the order has the same meaning across blocks. To reflect the within block design, one could use the following call in R

```
outlm <-aov(outcome~treat+Error(block/treat))
```

This gives the following ANOVA table.

```
   Error: block
             Df Sum Sq Mean Sq F value Pr(>F)
   Residuals  5  73.67   14.73
   Error: block:treat
         Df Sum Sq Mean Sq F value Pr(>F)
   treat      1  56.33   56.33    6.76 0.0482 *
   Residuals  5  41.67    8.33
   ---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

The above analysis indicates that there is indeed a significant difference between the treatments, after accounting for the block structure of the data.

Interaction plots can be drawn to visually evaluate the above results, and also to evaluate the results of the order analysis (if undertaken): (Fig. 12.4)

```
interaction.plot(block,treat,outcome)
interaction.plot(order,treat,outcome)
```

From the above plots, it would appear as though that there is a general increase in the mean response for both treatments as the block identifier increases. In regards

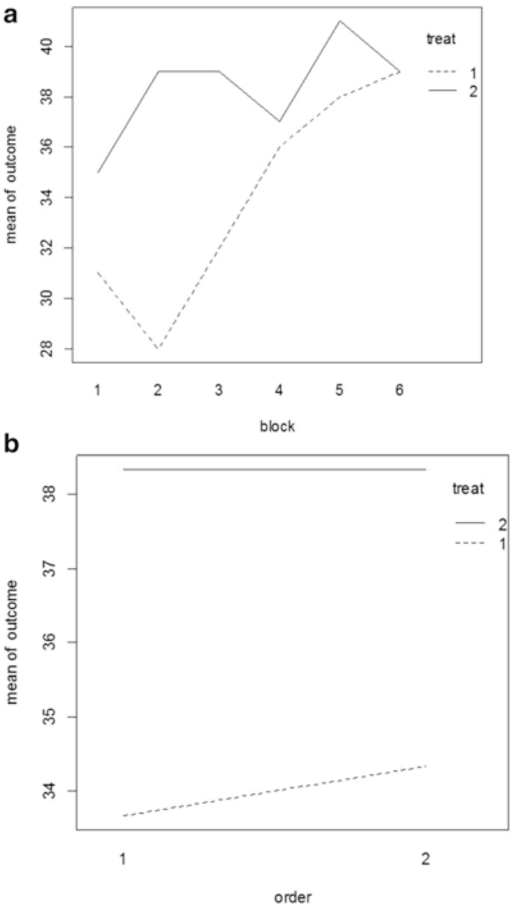


Fig. 12.4 Graphic representation of the effect of block randomization design on outcome (**a**) and of treatment order (**b**) on treatment outcomes. Both indicate possible carryover effects

to order, there is a slight suggestion that the effect of treatment 2 is increased when it is administered second. However, given the sample size of this study, this pattern may have been produced at random. An interaction between treatment and order could suggest that there is a carryover effect in treatment. Despite the typical assumption of a sufficiently long wash out period between treatments, the potential existence of a carryover effect should be investigated and either discounted or incorporated in the analysis.

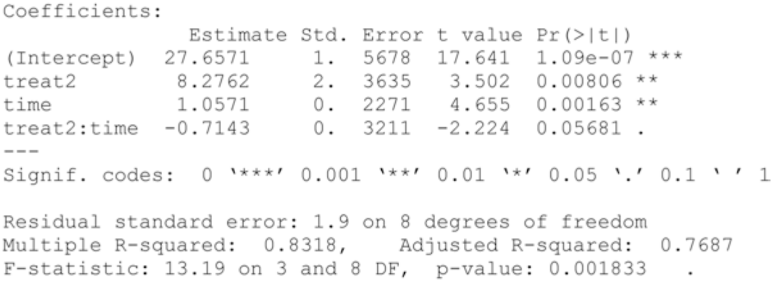
A linear modelling approach can be undertaken to test the effect of time on the outcome using the following command. Note that time, block, treatment and order can't all be included in the model, since there are insufficient observations to estimate all of these effects. In fact, order or block may be meaningless when time is included in the model, since they are both time factors, just defined differently.

For illustration, we omit blocks and order, and evaluate the effect of time alone. (Just for variety, we call the object containing the analysis in R a different name:

newout). Note that now there is no nesting in the model (since we are ignoring blocks and order), and an interaction between treatment and time is also estimated.

```
newout <-lm(outcome~treat*time)
summary(newout)
```

This yields the following results:



```
Coefficients:
               Estimate Std. Error t value Pr(>|t|)
(Intercept)    27.6571     1.  5678  17.641  1.09e-07 ***
treat2          8.2762     2.  3635   3.502  0.00806 **
time            1.0571     0.  2271   4.655  0.00163 **
treat2:time    -0.7143     0.  3211  -2.224  0.05681 .
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.9 on 8 degrees of freedom
Multiple R-squared:  0.8318,    Adjusted R-squared:  0.7687
F-statistic: 13.19 on 3 and 8 DF,  p-value: 0.001833   .
```

The estimated difference in treatment effects is much more significant and has actually increased when compared to our previous analysis. The sign of the interaction effect indicates that the increase over time is smaller in the treatment 2 group than in the treatment 1 group, which suggests that the difference between treatments decreases over the course of the trial. While the difference is 7.5 at time 1, it has completely disappeared by time 12. This suggests that the potential benefits of treatment 2 over treatment 1 are short lived. We note that in regards to time, the above model assumes that time is linearly related to outcome, depending upon treatment. It is important to check that this assumed linear relationship is reasonable. If not, then alternative parameterizations of time such as higher order polynomial terms or some grouping of time could be considered.

When fitting statistical models, it is important to check that statistical and modelling assumptions are appropriate. In the model described above, it is assumed that the residuals (difference between the observed and fitted data points) are normally distributed. Specifically, that the residuals are centered on zero, have constant variance, are independent and follow the bell shaped curve of the normal distribution. The following plots should be inspected to ensure that these assumptions are valid (Fig. 12.5).

```
par(mfrow=c(2,2))
plot(fitted.values(newout),residuals(newout),main="Fitted values
vs residuals")
plot(1:length(outcome),residuals(newout),type="l",main="Residu
als by order")
hist(residuals(newout))
qqnorm(residuals(newout))
```

From the 'Fitted values vs residuals' plot, it appears that the residuals are symmetrically distributed around zero, and there appears to be constant variability

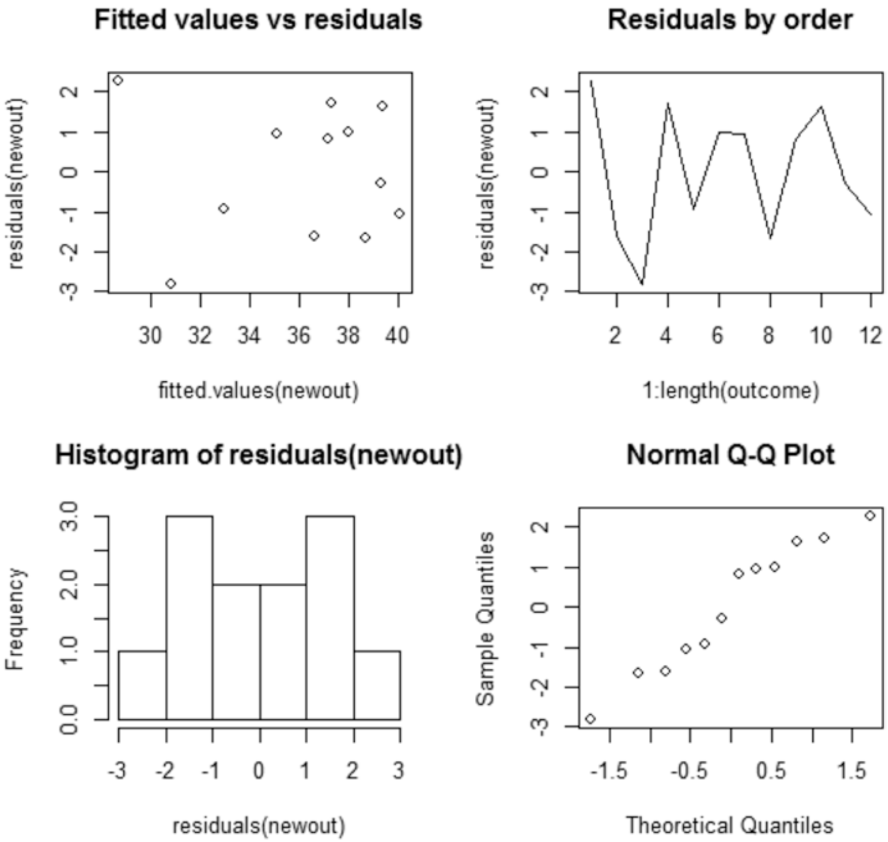


Fig. 12.5 Four means of assessing the normality of the example dataset

of these residuals among different fitted values. The 'Residuals by order' plot shows no patterns or runs in the residuals suggesting that independence is a valid assumption. The remaining two plots show that the characteristic shape of the normal distribution seems appropriate to describe the distribution of the residuals. When inspecting such plots, it is important to note that with such a small sample size, a variety of patterns could be observed, even if the residuals were generated from a normal distribution. Hence, some apparent anomalies may not necessarily void particular assumptions. For example, in this study, the histogram shows that the mode of the residuals is not 0. However, with such a small sample size, this is not enough to void the normality assumption.

It is also necessary to confirm other model assumptions. For example, it is assumed that each treatment group has equal variance and that the outcome has a linear relationship with time. These assumptions can be confirmed by inspecting the following plots (Fig. 12.6).

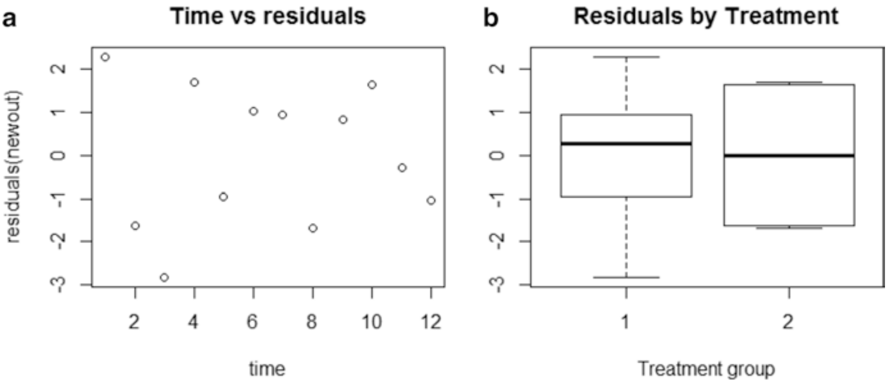


Fig. 12.6 Testing the effects of time (**a**) and mean and variability of each treatment group (**b**)

```
par(mfrow=c(1,2))
plot(time,residuals(newout),main="Time vs residuals")
boxplot(residuals(newout)~treat,xlab="Treatment group", main=
"Residuals by Treatment")
```

From the above ‘Time vs residuals’ plot, there is no evidence of a pattern suggesting that the relationship between the outcome and time is being captured appropriately. Further, from the above boxplot, there appears to be minor differences between the variability of the residuals by treatment group. However, this difference does not appear large enough to suggest actual differences in variability, particularly given the sample sizes in each group.

Despite the above ‘Residuals by order’ plot showing no reason to dispute the assumption of independence, measurements collected over time typically exhibit some form of autocorrelation. This feature of time series data reflects the dependence of a given outcome on data collected previously, and typically suggests that outcomes collected at times close to each other are more similar than outcomes collected further apart in time. There are many standard autocorrelation models for (stationary) time series data with the most common being the first-order autoregressive process. This model assumes that the covariance between errors is not zero but rather:

$$COV\left(\epsilon_t,\epsilon_{t+s}\right)=COV\left(\epsilon_t,\epsilon_{t-s}\right)=\sigma^2\rho^s,$$

where ρ^s is the correlation between errors that are s time units apart. We note that estimating this model involves estimating only a single additional parameter, ρ, and this model is only appropriate for continuous outcomes, that is, not discrete or ordinal outcomes. An initial estimate of this can be found as:

```
newout <- gls(outcome~treat*time)
cor(newout$res[-1],newout$res[-12])
[1] -0.2794359
```

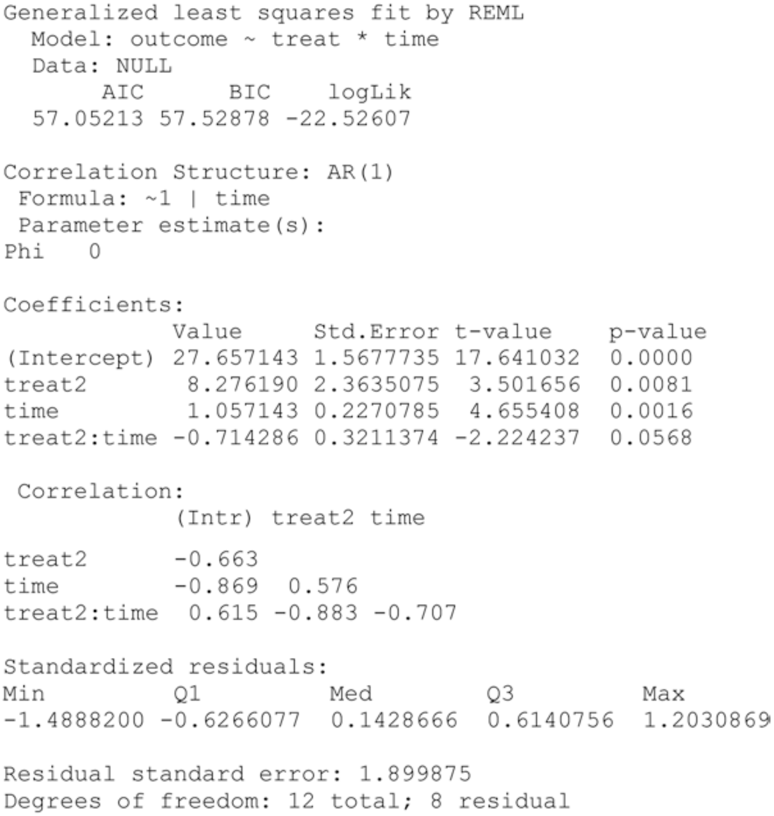
The correlation between errors at a variety of different *s* values can be explored by the acf function in R. That is,

```
acf(residuals(newout)
```

Models with such an error structure can be estimated under a generalized least squares framework via the gls function. This can be implemented in R as follows:

```
newout <-gls(outcome~treat*time,
correlation=corAR1(form=~1|time))
summary(newout)
```

This yields the following results:



```
Generalized least squares fit by REML
  Model: outcome ~ treat * time
  Data: NULL
       AIC      BIC    logLik
  57.05213 57.52878 -22.52607

Correlation Structure: AR(1)
 Formula: ~1 | time
 Parameter estimate(s):
Phi   0

Coefficients:
            Value     Std.Error t-value    p-value
(Intercept) 27.657143 1.5677735 17.641032  0.0000
treat2       8.276190 2.3635075  3.501656  0.0081
time         1.057143 0.2270785  4.655408  0.0016
treat2:time -0.714286 0.3211374 -2.224237  0.0568

 Correlation:
            (Intr) treat2 time

treat2      -0.663
time        -0.869  0.576
treat2:time  0.615 -0.883 -0.707

Standardized residuals:
Min          Q1          Med         Q3          Max
-1.4888200 -0.6266077  0.1428666  0.6140756  1.2030869

Residual standard error: 1.899875
Degrees of freedom: 12 total; 8 residual
```

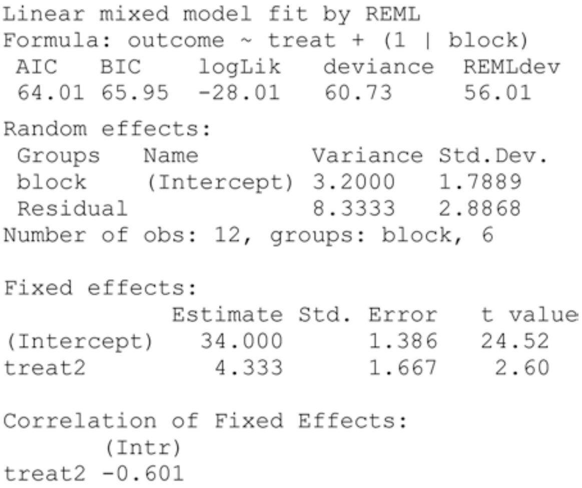
The autocorrelation parameter is estimated to be 0 suggesting this part of the assumed model is not needed. We note the difference between this estimate and our initial estimate of −0.28. This is due to the conditional and iterative nature of the generalized least squares estimation procedure. In general, the value of estimating this additional parameter can be investigated via a likelihood ratio test or by comparing values of different information criteria. Such criteria include the Akaike information criterion or AIC (Akaike 1974) and the Bayesian information criterion or BIC (Schwarz 1978).

Further analyses could be undertaken. For example, in the analyses thus far, we have considered the block effects to be fixed. Alternatively, one could consider such effects as random or more specifically random effects drawn from a population of block effects. This is useful in cases where the blocks considered in the experiment cannot be used again for data collection. In this study, blocks were used to reflect the variability in the response between different periods of time, hence this cannot be repeated exactly. In such cases, one is usually less concerned with the specific estimated block effects, but rather more interested in learning about the distribution of block effects. Having this understanding would be useful for experimentation into the future.

There are many different packages and functions used in R to estimate models with random effects (such models are also known as mixed effects models). Examples of packages include: lme4, nlme and asreml (with each using different function calls to estimate the mixed effects model). Below is an example of a mixed model involving blocks which can be estimated via the lmer function in the lme4 package.

```
library(lme4)
mixedmodelout <- lmer(outcome~treat + (1|block))
summary(mixedmodelout)
```

The output from the above call is given below:



```
Linear mixed model fit by REML
Formula: outcome ~ treat + (1 | block)
 AIC   BIC   logLik  deviance  REMLdev
 64.01 65.95 -28.01  60.73     56.01

Random effects:
 Groups   Name        Variance Std.Dev.
 block    (Intercept) 3.2000   1.7889
 Residual             8.3333   2.8868
Number of obs: 12, groups: block, 6

Fixed effects:
            Estimate Std. Error  t value
(Intercept)   34.000      1.386   24.52
treat2         4.333      1.667    2.60

Correlation of Fixed Effects:
       (Intr)
treat2 -0.601
```

From above, one can see that the estimated difference in treatment effects is the same between the fixed and mixed effects models. The output also shows that variability from two different sources (block and residual) was assumed. The usual assumptions regarding the residuals are made here. In regards to the block variability, it is assumed that the block effects are normally distributed around zero with a standard deviation estimated to be 1.8.

Discussion

This chapter has considered the statistical modelling and analytic aspects of N-of-1 trials. The preceding sections have provided illustrations of how statistical models and corresponding analyses can be developed for the more common designs encountered in N-of-1 trials. Other designs, perhaps involving different nesting of treatments, order and blocks, can be developed in a similar manner.

Mixed effects models were also discussed. Such models can be extended to account for a variety of factors whose effects can be considered as random draws from a population of effects. For example, this modelling approach can also account for the between subject variability of repeated measures data, and thus offers an approach to combining different N-of-1 trials conducted on many individuals (Zucker et al. 2010; Schmid and Duan 2014). Importantly, such a modelling approach can also handle sparse and/or unbalanced data that occur in studies for a variety of reasons including missing data.

The focus of this chapter has been on continuous response outcomes, that is, numerical response data. There are, of course, other data types such as binary or count data which could be measured from N-of-1 trials. In such cases, it is important to appropriately model the distribution of the response data. A wide variety of response data types (or distributions of data) can be modelled within a generalized linear modelling framework. Such models have three components: a distribution of the response, a linear predictor and a link function that relates the mean response to the linear predictor. It is the link function that appropriately re-scales the linear predictor to define different parameters in different distributions as a function of explanatory variables. These models can be implemented in R within the glm function for fixed effects models and in the glmer function for mixed effects models. In either case, one needs to specify the appropriate distribution of the data. For example, binary data could follow the Binomial distribution and count data could follow the Poisson distribution.

The importance of exploring the goodness-of-fit of all models considered cannot be understated. One part of this is assessing the appropriateness of all statistical and modelling assumptions. This exploration would focus on checking the validity of assumptions in regards to the distribution of the residuals, the appropriate inclusion of explanatory variables and the predictive ability of the particular model. As shown, the assumptions about the distribution of the residuals can be investigated via histograms, quantile-quantile plots, a plot of the residuals versus the fitted values and a plot of the residuals versus the order of the data. In terms of models that account for autocorrelation, the specified error structure needs to be checked. In the first-order autoregressive model implemented in this chapter, the exponential decay of ρ^s towards zero as s tends to infinity should be verified. In relation to checking the appropriate inclusion of explanatory variables, plots of the residuals versus each explanatory variable could be inspected for the presence of any patterns that are unaccounted for in the model. Other considerations for the mixed effects model include checking the appropriateness of the assumed distribution of the random effect estimates. Another part of assessing goodness-of-fit is investigating the accuracy and associated uncertainty of model predictions. These can be inspected visually or by cross-validation techniques such as leave-one-out procedures.

Further, we only very briefly touched on model choice. With such an array of potential models available for analysis and potentially many explanatory variables to be considered for inclusion into the model, it can be a difficult task to determine the most appropriate statistical model. This is a common statistical problem, and as such it is prevalent in the literature. A wide variety of the literature focuses on information criteria such as AIC and BIC. These criteria are calculated for each competing model, and are constructed in a manner that rewards goodness-of-fit but penalizes model complexity. This means more complex models are preferred or an additional explanatory variable is included into the model only if either of these choices significantly increases the goodness-of-fit of the model. We note that the choice between random effect models is difficult, in general. In such cases, one can consider the appropriateness of assuming a random effect and/or evaluate the variability of the random effect to determine the worth of inclusion. Other approaches based on the differences in deviance have also been considered in the literature.

The residual variance structure of proposed models requires careful consideration in relation to checking model assumptions and model choice. It can also relate to the sample size of the N-of-1 trial/s. In taking account of the trial design, we considered two forms of residual variance. The first was uncorrelated errors with a common variance parameter for all time periods and both treatments, and the second was a first-order autoregressive structure where errors were assumed to have a specific form of correlation depending upon distance apart in time. There are a variety of other choices that are worth considering in terms of model assumptions and benefit in regards to model choice. For example, one could consider a completely unstructured covariance matrix providing great flexibility is describing the covariance of the errors. Further, one could consider other autoregressive structures and/or different variance terms for each treatment. The benefits of assuming more complex variance structures, even if they actually exist, will ultimately be determined by the sample size of the study. That is, are there enough data points to actually observe the variance structure, and does assuming this structure significantly improve the goodness-of-fit of the model? Therefore, before trying a variety of different variance structures, one should consider which ones make sense given the study design (that is, the number of data points available to estimate the variance parameters).

The consideration of an appropriate variance structure is also relevant when a mixed effects model is being considered.

As the literature on N-of-1 trials grows, it will be interesting to critically evaluate the usage and utility of different analytic approaches. This is starting to emerge in meta-analyses of N-of-1 trials, where the reported results include descriptions of study design and analysis. The appeal of combining N-of-1 trials is being increasingly recognized. For example, Lillie et al. (2011) discuss motivations for, and examples of, such meta-analyses. They also identify a number of questions related to study design, namely randomization, carryover effects, washout periods and the use of blinding, baseline periods and placebo controls, and cite a number of articles that discuss the analysis of these trials (Kazdin 1982; Barlow and Hersen 1984; Spiegelhalter 1988; Rochon 1990). Meta-analytic methods for N-of-1 trials are described in a companion chapter, Chap. 16.

Conclusion

Appropriate modelling and analysis are crucial for accurate statistical inference and clinical decision support. However, they are only part of the larger picture: they depend critically on careful design and conduct of the study, and management and preparation of the data. Clear reporting of statistical methods, models and analyses, including the availability of code and data, will facilitate continual improvement in the way that these trials are designed, conducted, analyzed and used. This call for clarity and transparency in statistical analysis is not confined to this type of study, of course, but applies to all fields of quantitative scientific endeavor. It is hoped that this chapter provides some useful resources to assist in this challenge.

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom Control 19:716–723

Barlow DH, Hersen M (1984) Single case experimental designs: Strategies for studying behavior change, vol 2, 2nd edn. Pergamon Press, New York.

Kazdin AE (1982) Methods for clinical and applied settings, vol 368, Single-case research designs. Oxford University Press, New York

Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ (2011) The N-of-1 clinical trial: the ultimate strategy for individualizing medicine? Pers Med 8(2):161–173

Rochon J (1990) A statistical model for the "N-of-1" study. J Clin Epidemiol 43(5):499–508

Schmid CH, Duan N (2014) Chapter 4: The DEcIDE Methods Centre N-of-1 guidance panel statistical design and analytic consideration for N-of-1 trials. In: Kravitz RL, Duan N, The DEcIDE Methods Centre N-of-1 Guidance Panel (Duan N, Eslick L, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S) (eds) Design and implementation of N-of-1 Trials: a user's guide. AHRQ Publication No. 13(14)-EHC122-EF. Agency for Healthcare Research and Quality, Rockville, pp 33–53, Feb 2014. http://effectivehealthcare.ahrq.gov/index.cfm/search-for-guides-reviews-and-reports/?productid=1857&pageaction=displayproduct

Schwarz GE (1978) Estimating the dimension of a model. Ann Stat 6:461–464

Spiegelhalter DJ (1988) Statistical issues in studies of individual response. Scand J Gastroenterol Suppl 147:40–45

Zucker DR, Ruthazer R, Schmid CH (2010) Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. J Clin Epidemiol 63:1312–1323

Chapter 13
The Economics of N-of-1 Trials

Jennifer A. Whitty, Joshua M. Byrnes, and Paul A. Scuffham

Abstract This chapter focuses on the rationale, challenges and methodological considerations for evaluating the economics of N-of-1 trials. First, we outline the rationale for undertaking an economic evaluation alongside an N-of-1 trial, by describing two key economic questions that are likely to be of interest to researchers, policy makers and clinicians. Then we outline the methods for undertaking an economic evaluation, highlighting some methodological aspects that are of particular relevance for the economics of N-of-1 trials as opposed to more traditional clinical trials. Finally, we acknowledge that the economic evaluation of N-of-1 trials is still in its infancy. We reflect on the research agenda to further develop the potential for N-of-1 trials to inform optimal decision-making around treatment and the appropriate allocation of health care resources.

Keywords N-of-1 trials • Economics • Economic evaluation • Cost-effectiveness • Costs • Cost-effectiveness analysis • Benefits • Preferences • Economic methods • Efficiency • Heterogeneity • Decision-making • Resource allocation • Health technology assessment

J.A. Whitty (✉)
School of Pharmacy, The University of Queensland, Brisbane, Australia

Menzies Health Institute Queensland, Griffith University, Nathan, Australia

Centre for Applied Health Economics, School of Medicine, Griffith University, Nathan, Australia
e-mail: j.whitty@pharmacy.uq.edu.au

J.M. Byrnes • P.A. Scuffham
Menzies Health Institute Queensland, Griffith University, Nathan, Australia

Centre for Applied Health Economics, School of Medicine, Griffith University, Nathan, Australia
e-mail: j.byrnes@griffith.edu.au; p.scuffham@griffith.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_13

Introduction

In Chaps. 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, and 12 of this book, the authors have argued that N-of-1 trials provide an important research design with the potential to evaluate individual treatment effects and to estimate heterogeneity of treatment effects in a population (Gabler et al. 2011). However, the economics of N-of-1 trials needs to be considered also. The economics of N-of-1 trials relates to the economic value of undertaking a clinical N-of-1 trial as opposed to an alternative research study or usual practice (i.e. no trial). It also relates to the value of extending N-of-1 trials to compare the costs as well as benefits of two or more interventions at an individual level. These economic questions are, or should be, important considerations for health care funders, clinicians and patients. However, whilst conventional economic evaluation methods undertaken alongside traditional clinical trials, such as a randomized controlled trial, are well established, there are some nuances in the application of economic evaluation and interpretation of the findings in the N-of-1 trial context as opposed to the more traditional trial setting that need consideration.

Why Consider the Economics of N-of-1 Trials?

Imagine you are considering undertaking an N-of-1 trial to address the clinical question of whether a new intervention is superior to an existing intervention in either a specific patient, or at the aggregate level, in a heterogeneous cohort. There are two economic questions that might be of interest alongside this N-of-1 trial. The first relates to the outcome of the trial, and asks “Is the new intervention a cost-effective use of resources compared to an existing intervention, for the management of the specific indication in this specific patient (or alternatively in this cohort)?” In other words, does the new intervention provide acceptable value for money? The second question relates to the optimal research approach for answering this question: “Is an N-of-1 trial an economically viable research method to address this clinical question?” In this section, we outline approaches to each of these questions in turn. In the following section, we then outline some methodological considerations in addressing these questions.

Is the New Intervention a Cost-Effective Use of Resources?

In comparison to the relatively developed literature reflecting the use of N-of-1 trials to evaluate the clinical outcomes associated with an intervention (see for example the systematic review by Gabler and colleagues (2011)), economic evaluations undertaken alongside N-of-1 trials have been sparse. Karnon and Qizilbash (2001) were innovators in recognizing the potential for N-of-1 trials to provide estimates of individualized cost-effectiveness. Using a hypothetical example, they acknowledge the potential advantages of undertaking an economic N-of-1 trial and reflect on when it might be most beneficial to do so, alongside some considerations regarding the limitations of this approach. Subsequently, we are aware of only a small number of studies that have estimated cost-effectiveness alongside an N-of-1 trial, to assess whether a new intervention represents a cost-effective use of resources (Pope et al. 2004; Scuffham et al. 2010).

The Perspective of the Economic Decision

The economic decision in an N-of-1 trial is whether a new intervention is of acceptable cost-effectiveness compared with an existing intervention(s) (also referred to as the comparator). Cost-effectiveness is defined as the incremental cost of providing an additional unit of benefit, at the margin. When the analysis is at the individual level, this decision concerns the treatment of an individual patient. However, it is also feasible to undertake a cohort analysis from N-of-1 trial data to inform this decision at a population level. A cohort analysis in an N-of-1 trial achieves a similar aim to traditional clinical trials, but with the added benefit that much more refined data are available on the heterogeneity of the comparative costs and benefits in the population of interest. Therefore, in theory it should be possible to explore heterogeneity of the cost-effectiveness estimate across different individuals and to investigate the reasons for heterogeneity in the comparative costs and benefits associated with the intervention. Nevertheless, economic evaluations indicating the cost-effectiveness of interventions in N-of-1 trials thus far have not presented estimates at the individual level (Scuffham et al. 2010; Pope et al. 2004). Rather, they still aggregate results at the sample level and have not harnessed the unique potential of N-of-1 trials to describe heterogeneity within a sample. This might be at least in part because economics is conventionally of most use to inform the availability of a new intervention at a population level (i.e. should it be funded and for whom) than for treatment decisions at an individual level, for which the clinical effectiveness data and patient preferences are likely to take a primary role in decision-making.

Regardless of the level of analysis (individual or cohort) of N-of-1 data on costs and benefits, there are two main perspectives that are of interest in the evaluation:

1. The perspective of the public payer, i.e. should public funds be used to provide the new intervention for this specific patient or patient group? This perspective can be answered to some extent by conventional trials but only by assuming that the specific patient is an "average" member of the population or subgroup. It is answered to a much more refined degree by an N-of-1 trial, since an N-of-1 trial eliminates the between-subject variability in both the response to treatment and the associated cost of treatment. Costs in particular are usually subject to large variation across a cohort; thus elimination of this variation is statistically beneficial.
2. The perspective of the individual patient, i.e. is it worth an individual paying the additional cost for this new intervention out of their own pocket? This perspec-

tive cannot be strongly addressed in conventional trials, since we do not know a priori how an individual will respond to treatment, or their associated costs.

N-of-1 trials then have potential to provide data to support evaluation of the cost-effectiveness of new interventions and an assessment of who should receive them, and this information can potentially be used to inform individualized treatment decisions (Karnon and Qizilbash 2001).

Key Considerations with the Use of N-of-1 Trials to Evaluate Cost-effectiveness

Compared to traditional clinical trials, researchers have identified a number of specific considerations when assessing cost-effectiveness alongside an N-of-1 trial. First, N-of-1 trials are only applicable in a narrow range of conditions, as discussed in detail in Chap. 4. For an N-of-1 trial to be viable, the condition being treated will ideally be chronic and stable, the intervention will have a rapid onset of action, only a short carry over effect and will not be curative, and clinical response will be unpredictable at the individual level (Karnon and Qizilbash 2001; Scuffham et al. 2008b). In addition, an N-of-1 trial will be most beneficial in terms of optimizing value for money when the intervention under consideration is high cost, the cost differential between the intervention and comparator(s) is high, the proportion of responders is low, the proportion of non-responders continuing the intervention in the absence of a trial is high, and the cost of running the trial is low (Scuffham et al. 2008b). The N-of-1 trial is likely to offer economic benefits only in these particular circumstances. Thus, it is not well suited across all or even many clinical conditions.

As a consequence of their selective applicability, Karnon and Qizilbash (2001) raise ethical uncertainties around using cost-effectiveness estimates from N-of-1 trials to inform resource allocation at a population level. These concerns arise because the suitability of N-of-1 trials to only selected conditions means that alternative methods of economic evaluation (e.g. traditional clinical trials and economic modelling) must be used to assess the cost-effectiveness of interventions for the treatment of other conditions. This leads to inconsistency in the methodological approaches to evaluating the complete range of interventions for which public funding decisions must be made from the same bucket of funds. This inconsistency is a lesser ethical concern if individualized cost-effectiveness is used to inform individual treatment decisions in consultation with a patient. Moreover, where the recruitment or retention of trial participants would be difficult or impossible within a traditional randomized control trial, an N-of-1 trial still allows a randomized comparison of treatment to be made. Consequently, an N-of-1 trial in this circumstance might be justified.

One interesting point raised by Karnon and Qizilbash (2001) is the likely lack of power of N-of-1 trials. The small number of observations within any individual may not be sufficient to provide a statistically significant outcome at the individual analysis level, even though the outcome may be accepted as clinically decisive.

Pragmatically, it may not be feasible or acceptable to have a sufficient number of treatment episodes in one individual to test a statistical hypothesis. This lack of power for a clinical outcome will also likely apply to the individual assessment of costs and cost-effectiveness alongside the trial. Nevertheless, as highlighted by Karnon and Qizilbash (2001), a clinically definite outcome may be a sufficient basis on which to assess individual cost-effectiveness. Given a treatment decision must be made, and there is a cost associated with not accepting a new intervention that is optimal as well as to accepting one that is not optimal, the lack of statistical validity becomes "irrelevant" to the decision (Claxton 1999). It is generally accepted that it is appropriate for a cost-effectiveness analysis to proceed without the statistical significance of either the comparative costs or benefits or both. This is accepted on the basis that (i) the cost-effectiveness estimate is a ratio of the incremental cost and benefit and thus may be significant, even if each individual parameter is not; and (ii) sensitivity and scenario analyses can explore the level of uncertainty associated with the cost-effectiveness estimate.

It is also possible that the time horizon for the economic evaluation will need to be extended to capture the longer term costs and benefits associated with the intervention and comparator (Karnon and Qizilbash 2001); however, this is also the case with economic evaluation of traditional clinical trials.

Is an N-of-1 Trial an Economically Viable Approach to Inform Treatment Decisions?

The second economic question relates to whether an N-of-1 trial is an optimal approach to address a research question. In other words, is it worth undertaking an N-of-1 trial? This is a question that can be framed around the expected value of information provided by undertaking further very specific research. The answer relates strongly to the specific clinical or policy question that the trial is intended to address.

At the aggregate level, N-of-1 trials could be considered instead of (or alongside) traditional clinical trials to provide evidence of the comparative effectiveness and cost-effectiveness of an intervention. Here, the benefits of an N-of-1 trial are likely to be around the ability of N-of-1 trials to give a more refined insight into heterogeneity of response. The comparative costs and benefits of N-of-1 and traditional clinical trials in this regard are likely to be difficult to quantify economically and to our knowledge this has not been explored in the literature.

What is likely to be of substantially more interest is the economic viability of N-of-1 trials to use individual clinical response to inform the decision to continue a high cost intervention (Scuffham et al. 2008b, 2010; Kravitz et al. 2008). The potential of N-of-1 trials to provide a high level of individualized evidence to inform individual treatment decisions, based on individual response to treatment, gives them the capacity to be used to target access to high cost interventions (Scuffham et al. 2008b; Kravitz et al. 2008). In this case, the economic question is whether it is worth paying the cost of undertaking an N-of-1 trial to inform access decisions, as compared to "standard practice", where standard practice is no individualized N-of-1 trial. With standard practice, individuals could receive either an existing intervention (with possible suboptimal clinical outcome), or the new intervention (with possible unnecessary costs and adverse effects). However, they receive these without the targeting to response provided by data from an N-of-1 trial.

Scuffham and colleagues (2008b, 2010) showed the potential for use of N-of-1 trials to tailor the decision to continue a high cost intervention in healthcare decision-making. In the context of N-of-1 trials of celecoxib for osteoarthritis and gabapentin for chronic neuropathic pain, they reported fixed costs associated with undertaking an N-of-1 trial in the region of AU$23,000 with an additional variable cost of AU$1,300 per patient (expressed as 2003–2005 Australian dollar value). In a subsequent paper, including a third trial (of medications for the management of Attention Deficit Hyperactivity Disorder) with 1 year follow up, the estimated marginal cost of running an N-of-1 trial was reduced to AU$600 per patient (2006 Australian dollars) (Scuffham et al. 2010). The authors reported these costs to be partially offset by the savings generated in subsequent prescribing patterns (Scuffham et al. 2008b, 2010). Moreover, the health benefits gained from individualized treatment resulted in estimates for the incremental cost (AU$6,896 per life year or AU$29,550 per quality-adjusted life year gained) well within the range generally considered to provide acceptable value for money in Australia (Scuffham et al. 2008b; Harris et al. 2008). The authors concluded that the N-of-1 trial offers a realistic and viable option for improving access to selected high cost medicines in patients for whom management is uncertain (Scuffham et al. 2008b, 2010). However, despite this potential, the role of N-of-1 trials to target access to high cost interventions outside of the research setting has not yet been realized in practice. Nor, to our knowledge, has any funding mechanism been put in place to support the application of N-of-1 trials to address routine policy decisions.

Methods for Assessing Cost-Effectiveness Alongside an N-of-1 Trial

In this section we outline some methodological considerations when undertaking an economic evaluation in the N-of-1 trial context. We do not aim to outline the methods for undertaking an economic evaluation, which are detailed in other sources (the interested reader is referred, for example, to Drummond et al. 2005; Gold et al. 1996 for more detailed information on economic evaluation methods). Rather, we focus on the nuances of the application of economic evaluative methods and the interpretation of the findings in the context of N-of-1 trials as opposed to the more traditional clinical trial.

The Economic Question

It is important to define the relevant economic question which a cost-effectiveness analysis is to inform. For example, the question may be:

- "Is drug X cost-effective for the treatment of condition Y?"
- "Is drug X cost-effective when provided only to those who respond?"
- "Is conducting an N-of-1 trial cost-effective if used to identify those patients who respond to the test intervention, with the purpose of restricting prescribing to those who respond?"
- "What is the most cost-effective treatment for a specific patient?"

The primary difference between these questions is the perspective of the decision maker. The first is a question often asked of government or other funding agencies about newly developed technologies. In Australia and the United Kingdom for example, government funding for pharmaceuticals to be made available to their populations is based on the cost-effectiveness of that product. However, the cost-effectiveness analysis is often conducted with respect to the average outcome results for the cohort rather than outcome results for the individuals. That is, for some medications, some proportion of the patients may respond well to treatment but the efficacy outcome for responders is moderated when responders' outcomes are averaged with those who do not respond to the new treatment. N-of-1 trials can therefore be used to not only answer "is a particular drug cost-effective?" but also "is that drug cost-effective only for those who respond?"

N-of-1 trials can provide a scientifically robust approach to achieve the maximum potential outcome without the cost of ineffective treatment in those patients who don't respond. Consequently decision makers may include a response to treatment criteria for ongoing treatment. N-of-1 trials provide a tool to achieve that outcome. However, the cost of conducting N-of-1 trials to establish patient treatment must be incorporated into the analysis of the proposed new technology. For example, each patient who begins treatment would be enrolled into an N-of-1 trial. The cost of administering and analyzing the N-of-1 trial would then need to be included as part of the cost of the new medication when assessing the cost-effectiveness of that medication.

On the other hand, practitioners may wish to consider whether conducting N-of-1 trials is cost-effective in determining the optimal treatment for their patients. In this scenario, conducting an N-of-1 trial is compared to not conducting an N-of-1 in determining patient treatment. The alternative to an N-of-1 trial is typically an informal process of trial of treatment and monitoring. Additionally, practitioners may also wish to consider the cost-effectiveness of treatment in deciding on optimal treatment for each patient. That is, as opposed to only considering the outcome measure from each treatment option included within the N-of-1 trial, the cost-effectiveness of those treatment options may be included in determining the optimal treatment for that patient.

These alternative questions require different considerations in trial design. The first requires an N-of-1 trial with consideration of appropriate measurement of costs and outcomes for each patient averaged at a population level for a particular technology. The second question includes the potential benefit of only providing treatment to those who respond, the cost of conducting N-of-1 trials for every patient must also be considered. The third requires analysis of the cost and outcomes of patients who follow a formal N-of-1 trial procedure vs. the costs and outcomes of patients who adhere to an ad-hoc informal process for determining their treatment. The final question requires analysis of the cost-effectiveness of multiple competing treatment options at a patient specific level.

The Target Participants

Whilst evenness of randomization in terms of the baseline characteristics of participants is not a concern within N-of-1 trials given that all participants receive both active and control treatment, it remains important that the participants enrolled in the N-of-1 trials are representative of the population for whom treatment is being assessed.

For N-of-1 trials designed to inform whether or not a drug should be considered for funding, the trial population should be representative of the population that would seek reimbursement for that medication from the payer. This should be consistent with clinical guidelines for that condition and comparator.

If an N-of-1 trial is being compared to an alternative process for determining treatment, patients should be randomly assigned to undergo either the N-of-1 trial process or the alternative process. In this scenario distribution of trial participants between these competing processes should be even.

For N-of-1 trials designed to determine patient level optimal treatment or to inform a funding decision in an individual patient according to response, the target population will ultimately consist of patients for whom disease management is uncertain with a number of competing treatment options.

The Comparator

Within N-of-1 trials the comparator may either be current active treatment or placebo, with appropriate washout between treatments. N-of-1 trials require strict control of within participant blindness. In order to protect the cloak of blindness the following considerations should be made:

- Design of pharmaceutical pack;
- Equivalent size, color, smell of competing treatments;
- Treatment frequency.

Box 13.1: Selecting a Comparator

Economic evaluation always involves comparative analysis. For any analysis to be meaningful, selection of an appropriate comparator is essential. The National Institute for Health and Care Excellence (NICE) in the United Kingdom stipulates the use of the "best alternative practice" as the most appropriate comparator in an economic evaluation (NICE 2004). In contrast, the Pharmaceutical Benefits Advisory Committee (PBAC) in Australia require "the therapy likely to be most replaced by prescribers in practice" as the main comparator (Pharmaceutical Benefits Advisory Committee 2013).

See Box 13.1 for information on selection of an appropriate comparator.

One nuance of an N-of-1 trial is that the analysis uses the patient as their own control, and from an economic perspective this means evaluating the costs before and after the trial is possible. Therefore, there could be multiple comparators (including usual care, or the pre-trial treatment) in an N-of-1 trial (Scuffham et al. 2010).

Measuring and Comparing Costs

A critical component of any cost-effectiveness analysis (CEA) is appropriate identification, measurement and valuation of costs associated with the competing treatments. The first step is to define what resources need to be accounted for and how best to measure their use. The second step is to identify a per unit value for these resources from which to calculate a cost.

In order to identify all appropriate costs associated with each treatment option, first the perspective of the analysis must be established. For example, health care costs borne by one government agency would not affect the decision making of another government agency. Similarly, patient out of pocket costs may not be considered a cost from the perspective of a third party payer like a health insurance fund. Alternatively, one may wish to consider the costs from a societal perspective. That is all relevant costs across all parties.

Costs can be separated into either direct or indirect costs; Table 13.1 provides some example costs that might be considered. Direct medical costs are those related to providing medical services, such as a hospital stay, physician fees for outpatient visits, and drug costs (including the cost of the medication itself and any downstream adverse events). Direct nonmedical costs are those related to expenses, such as transportation costs, that are a direct result of the illness. Examples of indirect costs are lost time from work (absenteeism) and unpaid assistance from a family member.

Costs can further be categorized as fixed or variable costs. Fixed costs are those that aren't considered to change with the number of treatments provided.

Within an economic evaluation, it is important to focus on the difference in costs between the two competing alternatives. In the case of evaluating the cost-effectiveness of a pharmaceutical, costs for each patient and for each phase of the trial are collected. In an N-of-1 trial, the costs associated with the phase of the trial in which patients were on the proposed drug are compared with the costs associated with phase of the trial when patients were on the comparator.

However, in some circumstances a pharmaceutical may only be considered to be of acceptable cost-effectiveness if it is administered only to patients who respond to treatment. In this circumstance an N-of-1 trial can be an appropriate method for determining response to treatment. Funding agencies may specify that funding for treatment may continue in only those patients who have demonstrated response within an N-of-1 trial. Consequently, in order to assess the cost-effectiveness of a pharmaceutical product under these circumstances, the cost of conducting N-of-1 trials for each patient who begins treatment must also be considered within the evaluation.

Alternatively, the evaluation may seek to determine whether conducting N-of-1 trials are cost-effective compared to, for example, a trial and monitor approach. Subsequently, the costs of conducting an N-of-1 trial are compared to the costs of the trial and monitor approach.

At a patient level, CEA may be used to determine optimal treatment. In which case, whilst the cost of conducting an N-of-1 trial is not included, the costs associated with each treatment are compared for each patient. See Box 13.2 for information on accounting for time preferences: discounting future costs and outcomes.

Table 13.1 Examples of costs

Direct costs (health)	Direct (non-health)	Indirect
Practitioner	Transport	Time lost from work
Clinic/hospital	Telephone	Carers' cost and time
Administration		
Cost of the intervention and comparator		

Box 13.2: Accounting for Time Preferences: Discounting Future Costs and Outcomes

It is human nature to value those things gained or lost now more than things gained or lost in the future. The use of credit cards displays the phenomena of positive time preference perfectly. Many individuals prefer to have the enjoyment provided by an item purchased now and delay the cost of that purchase until later, despite the fact that it will cost more in the end when done this way (dependent on the interest rate attached to the card). Another example is smoking; smokers value the pleasure of smoking now more than the health losses resulting from smoking in the future (Cairns 1994).

(continued)

Box 13.2 (continued)

Discounting is the process of making current costs and benefits worth more than those occurring in the future to account for time preferences. An alternative way to think of this is that discounting is a process to convert expected future costs and benefits into present values. Although the discounting of health effects has raised some debate in the past (Torgerson 1999), it is now common practice to report both discounted and undiscounted results in analyses so decision makers can see the results of the analysis in terms of absolute costs and effects, regardless of when they occur, and the same results accounting for time preferences (Smith and Gravelle 2001).

A discount rate of 3.5 %, for both costs and effects, is recommended by NICE in the UK (NICE 2004). A discount rate of 5 % is recommended in Australia and the U.S. (Smith and Gravelle 2001). The impact on results of different discount rates should be included as part of the sensitivity analyses completed. With the discount rate selected (*r*), a calculation of the present value (*PV*) of a cost (or effect) incurred in (t) years, a future value (*FV*), can be made (1) (Table 13.2).

$$PV = \frac{FV}{(1+r)^t} \tag{1}$$

Table 13.2 Discounting a future cost of $1,000 and 1 Life Year (LY, effect) occurring in 5 years using different discount rates

Future value	Discount rate (%)	Present value
Cost ($)		
1,000	0.0	1,000.00
1,000	3.5	841.97
1,000	5.0	783.53
1,000	10.0	620.92
Effect (LYs)		
1	0.0	1.00
1	3.5	0.84
1	5.0	0.78
1	10.0	0.62

Comparative Health Outcomes

Clinical and summary health outcome measures are collected for each patient and for each phase of the trial, whether it is an N-of-1 trial or a traditional clinical trial. Health outcomes may be clinical measures (e.g. a frailty score, leukocyte count), natural units (e.g. visits to the family physician, days stay in hospital), or a patient reported outcome such as a pain or quality of life measure.

Subsequent to defining the appropriate research question, efficacy is then estimated either with respect to the proposed drug versus comparator, the proposed drug versus comparator for those who achieved pre-determined criteria of response, the N-of-1 trial vs. alternative treatment identification process, or the proposed drug versus alternative treatment options for each individual. The comparative approach adopted, in combination with the nature of the outcome measure, will determine the appropriate statistical analysis.

It is also important when undertaking an economic evaluation that all potential consequences are considered. Additional benefits associated with N-of-1 trials that are being proposed to identify optimal treatment for a patient or as part of the administration of a drug may include:

- Increased level of patient engagement with medication;
- Improved patient understanding of their disease;
- Greater treatment adherence;
- Higher levels of patient satisfaction.

The measurement and valuation of these benefits for individuals and responders (and non-responders) will be specific to each trial and setting but are a significant advantage of conducting N-of-1 trials.

Cost-Effectiveness Analysis

Cost-effectiveness analysis (CEA) is the main form of economic evaluation of healthcare interventions. Health outcomes are typically expressed as natural units (e.g. hospital admissions) or a clinical measure (e.g. HbA1c). Other forms of economic evaluation include

- Cost-utility analysis (CUA) where health outcomes are typically measured using utility weights and converted to quality-adjusted life years (QALYs),
- Cost-benefit analysis (CBA) where health outcomes are measured in the same units as costs (i.e. currency), and
- Cost-minimization analysis (CMA), which is useful where health outcomes between the intervention and comparator are equal.

Thus, CUA and CMA are special cases of CEA, where the outcome is typically measured as a comparative QALY gain, or is equal across treatments and so is not directly relevant to the cost-effectiveness estimate, respectively. CBA is conceptually different to CEA and is applied to a lesser extent than CEA, CUA or CMA in the health setting (Drummond et al. 2005). To our knowledge CBA has not been applied in the N-of-1 trial setting; therefore, we do not consider CBA further here. With the research question in hand, along with the costs and outcome, the next step in assessing cost-effectiveness is to complete an incremental analysis.

CEA Alongside a Traditional Clinical Trial

Incremental analysis in health economic evaluation involves calculation of the difference in cost, and outcomes, when one or more alternatives are compared to another (Schulman and Seils 2003). The cost of the intervention ($\mathrm{Cost}_{\mathrm{Int}}$) and the cost of the comparator ($\mathrm{Cost}_{\mathrm{Comp}}$) are calculated and used to provide an estimate of the incremental cost (13.1).

$$Incremental\ \ Cost = Cost_{Int} - Cost_{Comp} \tag{13.1}$$

Similarly, the difference between the effect of the intervention ($\mathrm{Effect}_{\mathrm{Int}}$) and the comparator ($\mathrm{Effect}_{\mathrm{Comp}}$) are calculated to provide an estimate of the incremental effect (13.2).

$$Incremental\ \ effect = Effect_{Int} - Effect_{Comp} \tag{13.2}$$

See Box 13.3 for information on statistical methods for incremental analysis of clinical trial data.

Box 13.3: Statistical Methods for Incremental Analysis of Clinical Trial Data

When decision analytic modelling is used, incremental analysis is an inherent part of the analytic process. When individual patient data are collected as part of a clinical trial, statistical techniques need to be implemented. It is important to remember that it is the mean differences in costs and health outcomes between the intervention group and the comparator group that are the important metrics for incremental analysis. Whilst the simplest of statistical technique is the t-test, cost data, and sometimes outcome data, are typically non-normally distributed with highly skewed distributions; thus commonly applied parametric tests such as the t-test are not usually appropriate for testing a difference between groups. The temptation may be to analyze geometric means or medians, to use non-parametric techniques such as the Mann-Whitney U test, or to perform some transformation on the data. None of these approaches is appropriate as they do not quantify and test for differences in the arithmetic mean of the raw cost; the arithmetic mean is the important summary statistic from both a social and budgetary perspective (Thompson and Barber 2000; Ramsey et al. 2005).

In addition, factors other than treatment allocation may explain differences in cost between patients. Multivariable linear regression may be used to evaluate costs after controlling for covariates; however, again the distributional properties of the data are often limiting. Non-parametric bootstrapping is one

(continued)

Box 13.2 (continued)

option to overcome distributional limitations, as is the application of generalized linear models (GLMs). GLMs can be used to address the specific distributional properties of the data as well as incorporating another additional feature of cost data that can be challenging, substantial zero costs in the data.

Given the challenging distributional properties of cost data obtained during clinical trials, it has been recommended that both simple univariate and multivariate analysis be used along with the presentation of different multivariate models so the uncertainty of results using different analytic techniques can be compared (Glick et al. 2007). The same advice applies to effect data.

For both cost and effect data, incremental analysis involves reporting of the mean difference in cost and the mean difference in effect along with measures of variability, precision and an indication of whether the observed differences are likely to have occurred by chance.

The final step in an incremental analysis is to produce an incremental cost-effectiveness ratio (ICER). This is a ratio of the difference in cost (incremental cost, ΔCost) and the difference in effect (incremental effect, ΔEffect) (13.3).

$$ICER = \frac{\Delta Cost}{\Delta Effect} \tag{13.3}$$

This produces the final metric of the analysis, the additional cost per unit of outcome.

Estimating CEA Alongside an N-of-1 Trial

A CEA alongside an N-of-1 trial (N-of-1 CEA) is constructed for each intervention arm for each patient. This differs to traditional clinical trials where the means for each intervention and comparator groups are used and the differences calculated. Where the trial design involves multiple cross-overs between treatment options, the incremental cost is estimated as the difference between the mean costs for each treatment for each individual (13.4).

$$Incremental\,Cost = \frac{\sum Cost_{\text{int}}}{n_{\text{int}}} - \frac{\sum Cost_{comp}}{n_{comp}} \tag{13.4}$$

In the case of differential periods of treatment, for example, 6 months of intervention, 6 months comparator followed by 3 months intervention and finally 3 months comparator, then the weighted average cost should be estimated reflecting the relative duration of treatment.

The incremental effect is estimated in a similar manner. The incremental cost-effectiveness ratio (ICER), that is the comparison of the difference in cost (ΔCost) and the difference in effect (ΔEffect) is then estimated for each individual patient.

$$ICER_i = \frac{\Delta Cost_i}{\Delta Effect_i} \tag{13.5}$$

Cost-Effectiveness Decision Criteria

Traditionally, the decision criteria for CEA is whether the cost per unit effect is below an accepted level (cost-effectiveness threshold) that the decision maker is willing to pay. Alternatively, for N-of-1 trials, the decision criteria revolves around each individual patient. As such the analysis is to determine whether for any one patient the incremental gains are achieved at an acceptable level of additional cost. It is then possible to summarize, for the entire cohort enrolled in the N-of-1 trial, the proportion of patients whose benefit was achieved at an acceptable cost.

Uncertainty

Of course, this description of incremental analysis provides only a point estimate of the final ICER. In any evaluation there will be a level of uncertainty in the estimates of both incremental cost and incremental effect and correspondingly in the final ICER. It is important to quantify this uncertainty.

For the evaluation of cost-effectiveness alongside a clinical trial, the mean and 95 % confidence interval (CI) provide information on the average patient response along with an indication of the uncertainty in this value. The same applies to ICER estimates.

For N-of-1 CEA, the uncertainty is with respect to the ICER estimate for each individual. From a single (or limited) cross over study design it is not possible to estimate a 95 % confidence interval or any measure of uncertainty. For example, a single cross-over design provides only one cost and one effect estimate for the comparative treatment (i.e. initial therapy) and one cost and one effect estimate for the intervention treatment (i.e., subsequent therapy post cross-over). Unless there are multiple crossovers per patient (e.g., at least five trials of the intervention with five trials of the comparator) it is unlikely that an N-of-1 trial will provide a sufficient number of data observations to estimate confidence limits for each individual in the trial.

Whilst ICERs undertaken alongside traditional clinical trials can be estimated using both parametric and non-parametric approaches, N-of-1 CEA are typically limited to non-parametric approaches such as bootstrapping. Non-parametric bootstrapping uses re-sampling with replacement from the given distribution from which to calculate a 95 % CI.

Sensitivity Analyses

After accounting for the uncertainty in the incremental cost, incremental effect and the final ICER using statistical methods, there remain additional uncertainties unrelated to sampling variation. This is the role of sensitivity analysis. Sensitivity analysis is the usual approach for CEAs alongside clinical trials; however these should be applied to N-of-1 CEA when comparing all costs and outcomes for the cohort, but are generally not appropriate at the individual level. For example, the price of a drug will be the same for all participants in an N-of-1 trial and cannot vary for each individual.

A one-way sensitivity analysis is used to evaluate the relative impact of the variables included in the analysis. Systematic variation of each variable across a plausible range of values, whilst holding all other variables constant, reveals the relative influence of each variable on the cost-effectiveness estimate (Briggs 1999). For example, the plausible cost of hospitalization may vary by 10 % of the baseline value used in the analysis. Re-running the analysis with the hospitalization cost reduced by 10 % and increased by 10 % will reveal the impact of this variation on the overall result for each individual. Completing multiple one-way analyses with a fixed variation (e.g. ±50 %) will reveal which variables have the greatest to the least impact on the overall result for all individuals within the trial. Alternatively, best and worst case values can be used. Another alternative form of multiple one-way sensitivity analysis is to complete a threshold sensitivity analysis. Instead of varying each parameter by a fixed amount, each parameter is varied to the extent where it changes the overall result of the evaluation. Threshold sensitivity analysis shows how much a particular variable needs to change for example, to result in the intervention under evaluation being no longer cost-effective for any patient or for all patients.

Regardless, it is important to justify the form of sensitivity analysis selected as well as the choice of parameters included and, for sensitivity analyses other than threshold, the range over which these parameters are varied (Husereau et al. 2013). The results of sensitivity analyses are best presented in a table and also diagrammatically, for example as a Tornado Diagram (Drummond et al. 2005).

Reflections on the Research Agenda

Economic evaluation alongside N-of-1 trials is in its infancy. We conclude this chapter by highlighting some of the key controversies and areas where future research is needed in order to progress the application of economic evaluation alongside N-of-1 trials and its relevance to decisions in practice.

Uptake of Economic Evaluation Alongside N-of-1 Trials by Healthcare Decision-Makers

Despite the potential of individual N-of-1 trials to optimize access to high cost interventions (Scuffham et al. 2008b), they do not appear to have been adopted to guide market access or subsidy decisions yet. Their role in this context is largely unexplored. Mixed methods research could investigate the feasibility and barriers to implementation (Kravitz et al. 2009). It would also be useful to gain insights into the relative merits of an individual level or cohort level analysis from N-of-1 trials as opposed to a traditional clinical trial, including the potential of N-of-1 trials to expand our understanding of the heterogeneity of treatment response and costs. N-of-1 trials offer patients an approach to effective and cost-effective individualized medicine. Identifying the intervention that the patient has the greatest response to enables the patient to make use of that intervention in the knowledge that it is effective for them. Moreover, this reduces wastage of scarce healthcare resources as the patient does not continue using an ineffective intervention for a protracted period until the clinician determines that there may be a better option; sometimes this can be months.

Measuring the Value of Patient Involvement

Evaluations to date have considered only the tangible clinical and cost benefits associated with N-of-1 trials. However, N-of-1 trials are closely aligned with the conceptual framework of patient-centered health care (McMillan et al. 2013). As such, they are likely to provide intangible benefits that go beyond direct health outcomes which are as yet not easy to measure. These benefits might include greater patient involvement with their care and decision-making, an improved clinician-patient relationship, and a more holistic understanding of the trade-offs and patient preferences around the use of high cost interventions (Karnon and Qizilbash 2001). Moreover, as a consequence of its individualized focus, it is quite possible that the N-of-1 research method might of itself produce improvements in health (Karnon and Qizilbash 2001; McMillan et al. 2013). Mechanisms for measuring and valuing the broader benefits for patients and society associated with the delivery of N-of-1 trials need to be explored.

Individual Patient Preferences

Arguably, the individualized nature of the N-of-1 trial and close involvement of patients in decision-making means the N-of-1 trial closely captures and accounts for individual patient treatment preferences in decision-making. This is aligned with previous attempts to describe improvements in health related quality of life at the individual level (Ruta et al. 1994). It is also consistent with current endeavors to establish methods for valuing the outcomes of health care using preference-based valuations of outcome at an individual level (Lancsar and Louviere 2008). However, the concept of optimizing individual treatment preferences from health care raises several interesting considerations relevant to the underlying theories of health economics. Conventional economic evaluation has evolved to evaluate "average" costs and benefits across a relevant cohort, and importantly to value "average" benefits based on the preferences of the general public (Scuffham et al. 2008a). It is argued that this approach to valuation is closest to the ideal of valuing benefits from behind a "veil of ignorance" (Rawls 1999), avoids potential bias due to patient self-interest, and allows the valuation to represent the preferences of all tax-payers who jointly bear the opportunity cost of a funding decision (Robinson and Parkin 2002). From a normative perspective, making individual funding decisions based on the valuation of individual patients as might occur in an N-of-1 trial is somewhat contradictory to this approach.

A related point has been raised by Karnon and Qizilbash (2001). If patients realize the outcome of the trial affects their treatment decision, and if they have a preference for a specific treatment, there is the potential for gaming. This may not be completely mitigated by strategies to avoid bias, such as randomization and blinding. How then do we control for patient preferences to avoid bias? One potential answer to this challenge might be borrowed from the literature on patient preference trials (Preference Collaborative Review Group 2008). Obtaining an indication of patient treatment preference before randomization, and controlling for this in any aggregate analysis, might mitigate any unintentional self-interest bias.

Nevertheless, despite these considerations relating to the use and valuation of individual treatment preferences, the benefits of understanding the range of responses both clinically and economically to a treatment is an important advantage of the N-of-1 trial approach. This adds complexity to the data available to make economic decisions, but may possibly add to the validity of the decisions made.

Conclusion

This chapter has outlined the rationale, challenges and methodological considerations for evaluating the economics of N-of-1 trials. The costs and effects for an individual in an N-of-1 trial are important to use to determine if the health benefits are sufficient to justify any additional ongoing costs. This chapter describes the approach to estimating the additional costs and health outcomes for individuals in an N-of-1 trial. The classification of costs, measurement of health outcomes, data transformations such as converting costs and outcomes to present values through discounting are described. The statistical methods for data analysis, and making decisions based on the comparative effectiveness and costs are presented. Finally, some reflections on the research agenda to progress the methods of economic evaluation alongside N-of-1 trials are outlined.

References

Briggs A (1999) Economics notes: handling uncertainty in economic evaluation. BMJ 319:120

Cairns J (1994) Valuing future benefits. Health Econ 3:221–229

Claxton K (1999) The irrelevance of inference: a decision-making approach to the stochastic evaluation of health care technologies. J Health Econ 18:341–364

Drummond MF, Sculpher MJ, Torrance GW, O'brien B, Stoddart GL (2005) Methods for the economic evaluation of health care programmes. Oxford University Press, New York

Gabler NB, Duan N, Vohra S, Kravitz RL (2011) N-of-1 trials in the medical literature: a systematic review. Med Care 49:761–768

Glick H, Doshi J, Sonnad S, Polsky D (2007) Economic evaluation in clinical trials. Oxford University Press, Oxford

Gold MR, Siegel JE, Russell LB, Weinstein MC (eds) (1996) Cost-effectiveness in health and medicine. Oxford University Press, New York

Harris AH, Hill SR, Chin G, LI JJ, Walkom E (2008) The role of value for money in public insurance coverage decisions for drugs in Australia: a retrospective analysis 1994–2004. Med Decis Making 28:713–722

Husereau D, Drummond M, Petrou S, Carswell C, Moher D, Greenberg D, Augustovski F, Briggs AH, Mauskopf J, Loder E, Force IHE, Force, I. H. E. E. P. G.-C. G. R. P. T. (2013) Consolidated Health Economic Evaluation Reporting Standards (CHEERS)–explanation and elaboration: a report of the ISPOR Health Economic Evaluation Publication Guidelines Good Reporting Practices Task Force. Value Health 16:231–250

Karnon J, Qizilbash N (2001) Economic evaluation alongside n-of-1 trials: getting closer to the margin. Health Econ 10:79–82

Kravitz RL, Duan N, White RH (2008) N-of-1 trials of expensive biological therapies: a third way? Arch Intern Med 168:1030–1033

Kravitz RL, Paterniti DA, Hay MC, Subramanian S, Dean DE, Weisner T, Vohra S, Duan N (2009) Marketing therapeutic precision: potential facilitators and barriers to adoption of n-of-1 trials. Contemp Clin Trials 30:436–445

Lancsar E, Louviere J (2008) Estimating individual level discrete choice models and welfare measures using best worst choice experiments and sequential best worst MNL. Centre for the Study of Choice, University of Technology, Sydney

Mcmillan SS, Kendall E, Sav A, King MA, Whitty JA, Kelly F, Wheeler AJ (2013) Patient-centered approaches to health care: a systematic review of randomized controlled trials. Med Care Res Rev 70:567–596

Nice (2004) Guide to the methods of technology appraisal. National Institute for Clinical Excellence (NICE), London

Pharmaceutical Benefits Advisory Committee (2013) Guidelines for preparing submissions to the Pharmaceutical Benefits Advisory Committee (Version 4.4). Australian Government Department of Health, Canberra

Pope JE, Prashker M, Anderson J (2004) The efficacy and cost effectiveness of N of 1 studies with diclofenac compared to standard treatment with nonsteroidal antiinflammatory drugs in osteoarthritis. J Rheumatol 31:140–149

Preference Collaborative Review Group (2008) Patients' preferences within randomised trials: systematic review and patient level meta-analysis. BMJ 337:a1864. doi:10.1136/bmj.a1864

Ramsey S, Willke R, Briggs A, Brown R, Buxton M, Chawla A, Cook J, Glick H, Liljas B, Petitti D, Reed S (2005) Best practices for economic evaluation alongside clinical trials: an ISPOR RCT-CEA task force report. Value Health 8:521–533

Rawls J (1999) A theory of justice. Belknap Press of Harvard University Press, Cambridge, MA

Robinson A, Parkin D (2002) Recognising diversity in public preferences: the use of preference sub-groups in cost-effectiveness analysis. A response to Sculpher and Gafni. Health Econ 11:649–651

Ruta DA, Garratt AM, Leng M, Russell IT, Macdonald LM (1994) A new approach to the measurement of quality of life. The patient-generated index. Med Care 32:1109–1126

Schulman K, Seils D (2003) Clinical economics. In: Max M, Lynn J (eds) Interactive textbook on clinical symptom research. National Institutes of Health, Bethesda

Scuffham PA, Whitty JA, Mitchell A, Viney R (2008a) The use of QALY weights for QALY calculations: a review of industry submissions requesting listing on the Australian Pharmaceutical Benefits Scheme 2002-4. Pharmacoeconomics 26:297–310

Scuffham PA, Yelland MJ, Nikles J, Pietrzak E, Wilkinson D (2008b) Are N-of-1 trials an economically viable option to improve access to selected high cost medications? The Australian experience. Value Health 11:97–109

Scuffham PA, Nikles J, Mitchell GK, Yelland MJ, Vine N, Poulos CJ, Pillans PI, Bashford G, Del Mar C, Schluter PJ, Glasziou P (2010) Using N-of-1 trials to improve patient management and save costs. J Gen Intern Med 25:906–913

Smith DH, Gravelle H (2001) The practice of discounting in economic evaluations of healthcare interventions. Int J Technol Assess Health Care 17:236–243

Thompson S, Barber J (2000) How should cost data in pragmatic randomised controlled trials be analysed? BMJ 320:1197–20

Torgerson D (1999) Discounting. BMJ 319:914–915

Chapter 14
Reporting N-of-1 Trials to Professional Audiences

Margaret Sampson, Larissa Shamseer, and Sunita Vohra

Abstract Whether an N-of-1 trial is undertaken to inform a particular clinical decision or to test a hypothesis, publishing it in the professional literature may inform other clinical decisions and contribute to the research evidence base. A well-reported N-of-1 trial will provide the transparency needed for readers to critically appraise the work and determine if it is applicable to their situation. A well reported trial can be replicated and, once replicated, results can be aggregated to provide stronger and more compelling evidence. This chapter will consider how to describe the individual and aggregated data of N-of-1 trials for professional audiences. It describes in detail a reporting guideline for N-of-1 trials, CENT (**C**onsort **E**xtension for reporting **N**-of-1 **T**rials). CENT provides a structured format to ensure that the main journal report is sufficiently detailed that it can be critically appraised and replicated. As well, prospective registration of the trial and data deposit is discussed as means to further increase the transparency and completeness of reporting.

Keywords N-of-1 trials • Reporting guideline • CENT • CONSORT • Checklist • Transparency • Replication • Publishing • Protocol registration • Data deposit

M. Sampson (✉)
Children's Hospital of Eastern Ontario, Ottawa, Canada
e-mail: msampson@cheo.on.ca

L. Shamseer
Ottawa Hospital Research Institute, Ottawa, Canada
e-mail: lshamseer@ohri.ca

S. Vohra
Department of Pediatrics, University of Alberta, Edmonton, Canada
e-mail: svohra@ualberta.ca

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_14

Reporting Standards

Since the early 1990s, there have been concerted efforts to improve the completeness and clarity of reporting of healthcare research (Moher 2009). Philosophically, making trial results available to a professional audience has been stated as an ethical imperative. The Declaration of Helsinki states that

> Researchers have a duty to make publicly available the results of their research on human subjects and are accountable for the completeness and accuracy of their reports. All parties should adhere to accepted guidelines for ethical reporting. Negative and inconclusive as well as positive results must be published or otherwise made publicly available. Sources of funding, institutional affiliations and conflicts of interest must be declared in the publication. Reports of research not in accordance with the principles of this Declaration should not be accepted for publication. Paragraph 36 (World Medical Association General Assembly 2013)

This requirement would clearly apply to N-of-1 trials with a research focus, but many have a clinical focus, being designed to inform particular clinical decisions. It seems likely that many of these clinically oriented trials are not shared in the professional literature (Price and Grimley Evans 2002). However, even in these cases, the Declaration encourages evaluation of safety and efficacy and recommends that information be recorded and made publicly available (Paragraph 37).

Positive and negative results are equally important – reporting only successful trials, or only the successful outcome measures within trials, will result in a distorted picture of the efficacy of the intervention. Publication bias, that is, selective reporting of positive trials, has been extensively documented in the randomized controlled trial literature (Dwan et al. 2013); similar problems are likely to exist across the spectrum of health research, including in N-of-1 trials, perhaps to a larger degree due to fewer regulations and less monitoring. Increasingly, there is pressure from consumer, legal and professional sources to register trials and publicly report their results (Goldacre 2014; Lefebvre et al. 2013). Reporting guidelines assist in this. A reporting guideline is typically a consensus-based document, which provides authors with a minimum set of information that should be completely reported for a particular research design or design aspect (Moher et al. 2011). Importantly, reporting guidelines are not a judgment on the quality of the research (Moher 2009) although transparency in reporting by authors enables readers to better gauge the quality of the conduct and design of reported/published research.

Since the publication of the first scientific article around 1665, the overall organization of articles has become more formal and less literary in style and, in the twentieth century, the IMRAD format of Introduction, Methods, Results and Discussion has been adopted in medicine, becoming dominant by approximately 1965 (Sollaci 2004). The narrative or chronological approach persisted in abstracts until 1987 when the Annals of Internal Medicine introduced the structured abstract, but only for clinical trials (Huth 1987), with the aim of assisting readers in quickly judging the applicability and validity of findings of an article to clinical practice

(Haynes et al. 1990). By 1993 the International Committee of Medical Journal Editors recommended the use of structured abstracts and, by 1995, nearly three-quarters of clinical trial reports listed in MEDLINE used that format (Nakayama et al. 2005). Structured abstracts are also known as “more informative abstracts” – MIAs (Haynes et al. 1990).

In keeping with the early intent of making these reports of trials easier to find (Haynes et al. 1990), studies sponsored by the National Library of Medicine in 1995 and 2011 demonstrated that articles that had structured abstracts indeed had more retrieval points (MeSH terms and text words) than MEDLINE records as a whole (Ripple et al. 2011). While studies considering only searchability by study design did not find that structured abstracts helped (Wilczynski et al. 1995; Stevenson and Harrison 2009), such abstracts have been shown to improve reporting of study population, intervention and outcomes (Sharma and Harrison 2006), the elements that form the basis of structured clinical questions used to guide searches of the professional literature (Schardt et al. 2007).

CONSORT, CONSORT Extensions and Their Impact

Reporting guidelines are formalized extensions of these early efforts of IMRAD and structured abstracts. Among the first guidelines was a proposal for the structured reporting of randomized trials in 1994 which quickly evolved into the Consolidated Standards of Reporting Trials (CONSORT) Statement in 1996, focusing on the reporting of parallel group randomized controlled trials (Begg et al. 1996). It has been revised twice, in 2001 (Moher et al. 2001, 2010a; Schulz et al. 2010) adopted by over 600 biomedical journals and cited over 10,000 times. Reporting guidelines for other study designs soon followed, as did the adaptation of CONSORT to different types and aspects of trials; harms (Ioannidis et al. 2004) and non-pharmaceutical interventions (Boutron et al. 2008b) as examples. Over 200 reporting guidelines for a variety of research designs and types can be found in the EQUATOR (Enhanced Quality and Transparency of Reporting) Network Library (www.equator-network.org).

In 2010, guidance for the developer of reporting guidelines was published by the EQUATOR Network (Moher et al. 2010b). Its process was used to develop reporting guidance for N-of-1 trials and is described below.

In 2014, there were reporting standards that address particular study designs (e.g. randomized controlled trials, cohort studies, diagnostic studies, case reports), study type (e.g. quality improvement), or type of data (e.g. harms), with at least 90 guidelines or standards in existence (Moher 2009). Reporting standards discussed in the chapter are shown in Table 14.1. Research evidence has demonstrated that reporting standards do indeed increase the quality of reporting (Turner et al. 2012; Wen et al. 2008).

Table 14.1 Reporting guidelines abbreviations

CENT	CONSORT Extension for N-of-1 Trials (Vohra et al. 2015)
CONSORT	CONsolidated Standards Of Reporting Trials – revised (Schulz et al. 2010)
CONSORT extensions[a]	CONSORT for abstracts (Hopewell et al. 2008)
	CONSORT for harms (Ioannidis et al. 2004)
	CONSORT for non-pharmacologic treatment interventions (Boutron et al. 2008a)
	CONSORT PRO reporting of patient-reported outcomes in randomized trials (Calvert et al. 2013)
EQUATOR network	Enhancing the QUAlity and Transparency Of health Research (Morris 2008)
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses (Moher et al. 2009)
STRICTA	STandards for Reporting Interventions in Clinical Trials of Acupuncture – revised (MacPherson et al. 2010)
TIDieR	Template for Intervention Description and Replication (TIDieR) checklist and guide (Hoffmann et al. 2014)

[a]See the CONSORT web site at http://www.consort-statement.org for additional guidelines. This is a partial list and includes only reporting guidelines mentioned in this chapter

Evolution of the CONSORT Extension for N-of-1 Trials (CENT)

In 2006 a team led by Sunita Vohra of the University of Alberta, Canada, set out to examine all published reports of N-of-1 trials. Our objective was to ascertain the range of designs used to conduct N-of-1 trials; methods used in trial statistical analysis; and methods for combining data from a series of N-of-1 trials. While we later scaled back our efforts to focus on ABAB designs, we initially examined AB, ABA and ABAB approaches. What became clear quite early on, and was crystal clear after reading several hundred reports, was that it was very often difficult to tell what the investigators had done. We wanted to study the number of periods and pairs, the use of blinding, run-in and washout periods. We wanted to know the number of measurements per period, stopping criteria, types of outcome measures (e.g. subjective vs. objective, validated population-specific measures vs. patient- and symptom-specific measures) and methods for adverse event reporting. It was often not clear what the diagnosis was, and whether there were any concurrent conditions or therapies. Often, the description was inadequate to assess if the comparator used was appropriate. One frequently had to read and re-read, looking for clues as to how many treatment periods were administered, in what sequence and how that sequence was determined. Partly because many of the reports were so difficult to unravel, we eventually realized that new studies were being published faster than we could screen them and extract the data.

Over a period of years, we concluded that a standardized method of reporting of N-of-1 trials, adapted from the CONSORT statement for randomized controlled trial (RCT) reports, could help investigators improve the quality and consistency of their N-of-1 trial reports. Based on this preliminary work, and with funding support from a variety of agencies including the Canadian Institutes of Health Research, we developed the CONSORT Extension for N-of-1 Trials or CENT for short.

The Development of CENT

Following the process recommended by leaders of the EQUATOR group, (Moher et al. 2010b) we continued the systematic review of design and reporting of N-of-1 trials. We formalized an international steering committee of members with wide-ranging experience in clinical trial methodology and reporting guideline development to lead the development of CENT. Members were Doug Altman, Nick Barrowman, Cecilia Bukutu, Gordon Guyatt, David Moher, Margaret Sampson, Robyn Tate, Larissa Shamseer and Sunita Vohra. We then went through a consensus building process, engaging a larger group of international experts, to determine which items were most important to report, at a minimum, in reports of N-of-1 trials. This list of candidate items was drawn from our systematic reviews and from the items that make up the 2010 CONSORT checklist (Moher et al. 2011). The experts who participated were 56 individuals whose expertise variously included N-of-1 trials, biostatistics, clinical epidemiology and reporting guideline development. Biomedical journal editors and health research funders were included in that group, as were practicing physicians who had been identified as having been involved in N-of-1 trials. The consensus building included a two-round modified Delphi process in which these experts were surveyed to reduce the list of candidate items, setting aside those considered less important (Jones and Hunter 1995). This was followed by a consensus meeting to finalize the essential concepts to be included. After the concepts were finalized, the steering group organized and refined the concepts into checklist items. This checklist was first approved by meeting participants then circulated to solicit feedback from those invited to the meeting but unable to attend. The details of this process are fully enumerated in the CENT statement (Vohra et al. 2015). The steering committee then developed a supporting document providing explanation and elaboration, with examples of good reporting, for each item on the checklist. This document is commonly known as the "E&E" (Shamseer et al. 2015).

How to Use CENT

There are two main tools to support those using the CENT guidelines. The first is a checklist of 25 items, some with sub-elements that were determined to be the essential aspects of an N-of-1 trial to be reported. The checklist follows the IMRAD structure, so can be thought of as a detailed document outline. The checklist is available as part of the CENT statement (Vohra et al. 2015). The accompanying E&E gives the rationale for reporting each item (Shamseer et al. 2015). In some cases, such as randomization and blinding, the purpose of describing the research procedures is to allow the assessment of observer bias. In other cases, the purpose is to avoid bias in the interpretation of results, for example, by stating whether carryover effect or period effect was assessed. Some elements are needed so that meta-analysis, or aggregation across trials, is possible – these include reporting of measures of error and precision. Whatever the reason for inclusion in the checklist, the E&E explains the rationale and provides examples from published accounts of N-of-1 trials.

While most elements of the CENT checklist are intended for reports of both individual N-of-1 trials and series of N-of-1 trials, there is need for some nuanced reporting differences for some checklist items between these two types of reports and, where applicable to both individual and aggregated reports, examples are given of each type.

Fundamentally, detailed reporting of key elements of the methods and results for N-of-1 trials enables the reader's assessment of the validity of the research and both the clinician's and researcher's optimal use of N-of-1 trial findings, either in clinical care or future research.

Although reporting guidelines are not intended to dictate how a study should be designed and conducted, full reporting is easiest with forethought and planning. Thus, these two elements, the CENT checklist and the E&E, are helpful to the researcher in both protocol development and manuscript preparation once a trial (or series of trials) is complete. The CENT guidance can be used by journal editors and peer reviewers to assess the merits of a research report and request that gaps be filled. Many journals that have endorsed the CONSORT checklist require that authors provide a completed checklist, indicating the page number of each item in the manuscript when the manuscript is submitted for consideration (e.g. JAMA Instructions For Authors (JAMA 2014)). Finally, post-publication, the user of the published article can critically appraise the work.

The rest of this chapter addresses how specific aspects of the N-of-1 trial should be reported and will draw heavily from the CENT statement (Vohra et al. 2015) and the CENT E&E (Shamseer et al. 2015). The reader may wish to keep the CENT checklist at hand while reading. All CONSORT extension checklists can be downloaded from the EQUATOR web site (http://www.equator-network.org/).

Major Elements of Reporting

Reports of healthcare research usually follow the IMRAD format of introduction, methods, results and discussion, regardless of study design. Reporting guidelines can be thought of as an expanded table of contents of a report, with various subsections tailored to the design being reported. CONSORT is designed to optimize the reporting of parallel group trials and the CENT extension adds elements specific to the N-of-1 design.

Writing the Title and Abstract

Main elements within CENT that differ from CONSORT begin with Item 1a, the title of the manuscript – it should identify as an "N-of-1 trial" in the title, and for a series, identify the design as "a series of N-of-1 trials". The abstract should be a structured summary of trial design, methods, results and conclusions. Detailed guidance is available in Table 14.1 of the CENT E&E (Shamseer et al. 2015) as well as the CONSORT extension for abstracts, designed to cover both abstracts of journal articles and conference abstracts (Hopewell et al. 2008).

Writing the Introduction

Authors should state the scientific background, explain the rationale, including the rationale for using the N-of-1 design and state the specific objectives or hypotheses of the trial. It may be helpful to clarify whether the trial was done as research or as clinical care (Punja et al. 2014).

Writing the Methods Section

As would be expected, there are substantial differences in the methods section of an N-of-1 trial and a parallel group trial. In both cases, this section should describe the study design, the participants, interventions and outcomes.

Trial Design

For N-of-1 trials, authors would describe the planned number of periods and the duration of each period (including run-in and wash out, if applicable) with rationale. In addition, if the report describes a series of trials, authors should state whether and how the design was individualized to each participant, along with an explanation of the series design.

Throughout the report, and beginning with the trial design, any deviation from the planned design, such as a change in number or length of periods, should be described and explained. The reasons for the deviation may be important in interpreting the results.

Participant(s)

A description of the study participants is needed; readers should be able to clearly understand the diagnosis or disorder, the diagnostic criteria used and any co-morbid conditions and concurrent therapies. For a series, an additional description of the eligibility criteria for trial participation should be included; however the description of the actual trial participants should be reported in the results section. As well, the methods section should include a description of the settings and locations where data were collected and the dates defining the periods of recruitment and follow-up.

Interventions and Outcomes

The interventions for each period should be described in sufficient detail to allow replication. The description should include how and when the interventions were actually administered. A strength of N-of-1 trials is that interventions can generally be tailored to meet a patient's unique profile, (Guyatt et al. 2000) and so the intervention as tested needs to be fully and clearly described. Several other CONSORT extensions are available that can guide reporting of herbal interventions (Gagnier et al. 2006), acupuncture (MacPherson et al. 2010) and non-pharmacological treatments (Boutron et al. 2008b). In addition, detailed guidance on effectively describing interventions can be found in the TIDieR – the Template for Intervention Description and Replication Checklist and Guide (Hoffmann et al. 2014).

Measurement properties, that is, validity and reliability, of outcome assessment tools are needed. Any changes made to the selection of trial outcomes and measurement instruments after the trial commenced should be stated and reasons for the change explained. Supplemental guidance on patient-reported outcomes is available through CONSORT PRO (Calvert et al. 2013).

Population, intervention, comparison and outcome are the classic elements of a clinical query. At this stage of the report preparation, authors should reflect on whether they have given a sufficiently clear description of these elements so that the report can be found by a search of those essential parameters. Although sources such as PubMed do not allow readers to search the full text of articles, a good description will enable indexers to assign useful subject headings and thereby make the article easier to find.

Sample Size, Randomization, Blinding

Also included in the methods section are the more technical elements of the design – sample size, allocation concealment, randomization, blinding and statistical methods – all of which need to be described in enough detail to permit critical appraisal and replication.

In discussing these elements, it is helpful to keep in mind that reporting guidelines address reporting – they are not prescriptive regarding how a trial should be conducted, therefore they do not mandate that a trial should or should not be randomized or blinded. CENT takes no position on whether statistical analysis is appropriate for N-of-1 designs. However, for each of these aspects, the report should make clear what was done. In N-of-1 trials, randomization refers to the random assignment of a patient to a treatment within a pair or block of a pre-specified size. Thus, for randomization, the reader needs to know whether the order of treatment periods was randomized and, if so, the method used to generate allocation sequence. Equally, if a counterbalanced design is employed so that treatment order (e.g., AB or BA) is systematically alternated (e.g., ABBA or BAAB), this should be reported. Whatever approach is selected, the mechanism used to determine the order of treatments should be described by authors, along with the rationale.

Further, when applicable, the mechanism used to generate the randomization sequence should be described in enough detail to enable readers to gauge whether the method used was robust. The unit of randomization, such as within a pair or block, or if treatments were simply alternated after randomly assigning the starting treatment, should be reported. If blocking was used, the block size should be reported as well as whether the size was fixed or randomly decided.

Following the description of how the sequence of periods was determined, the full intended sequence of periods needs to be stated. For series of N-of-1 trials, where the sequence is different for each individual trial, it may not be possible to report the planned sequence for each trial in the text. Sequences for each individual trial may instead be included as an appendix.

While there are considerable differences in randomization between parallel group and N-of-1 trials, allocation concealment and blinding are similar. Allocation concealment, that is, any steps taken to conceal the sequence until interventions were assigned, should be described. As part of this description, authors may need to describe the mechanism used to implement the sequence (such as sequentially numbered containers), who generated the sequence, who enrolled participants and who assigned participants to interventions. Allocation concealment may be one of the more poorly understood aspects of trial design and conduct and is often confused with randomization or blinding. Interestingly, a recent systematic review looking at completeness of reporting of RCTs found that allocation concealment was reported adequately twice as often in RCTs in CONSORT-endorsing journals than in non-endorsing journals. This was the biggest gain of the 27 outcomes assessed (Turner et al. 2012). It may be useful at this point to remember that the E&E provides examples of real life reporting for all elements in the checklist, including allocation concealment (Shamseer et al. 2015).

As with randomization, blinding may not occur in all N-of-1 trials. But, if blinding was used, it should be clear who was blinded after assignment to interventions (for example, participants, care providers, those assessing outcomes) and how this was done. If relevant, authors should describe the similarity of the intervention under investigation and the control or comparison condition.

Analytic Methods

Analytic methods for N-of-1 may include two broad types of approaches: visual analysis and statistical analysis. There is some difference of opinion as to which approach is preferable, thus authors should state which approach to analysis they used (if not both) and the reasons. In line with recommendations made by the International Committee for Medical Journal Editors (ICMJE) and the CONSORT group, analytical methods should be described "with enough detail to enable a knowledgeable reader with access to the original data to verify the reported results" (International Committee of Medical Journal Editors 1997).

Many N-of-1 trial authors provide a visual representation of the data which allows readers to inspect the slope, variability and patterns of the data and the overall reliability and consistency of treatment effects (Gage and Lewis 2013; Horner et al. 2012). Analytic aids such as a line of best fit may sometimes be used to facilitate interpretation of visually presented data. If such analysis was done, authors should describe how and why it was carried out.

For a series of N-of-1 trials, methods of any quantitative synthesis of individual trial data should be described fully, including subgroup analyses, adjusted analyses and how heterogeneity between participants was assessed. Authors may find it helpful to consult the PRISMA Statement for specific guidance on reporting syntheses of multiple trials (Moher et al. 2009).

Ethics

Finally, authors should include a statement in the methods section about the research ethics status of the N-of-1 trial. If the N-of-1 trial was undertaken solely to better manage an individual's treatment, i.e. as a form of enhanced clinical assessment, then the trial may not require institutional ethics review board oversight (Punja et al. 2014; Mahon et al. 1995; Irwig et al. 1995). Whether the report represents an undertaking under the auspices of research or clinical care should be made clear and if it is research, the report should cite the institutional ethics board that reviewed and approved the research study (Punja et al. 2014).

Writing the Results Section

Moving on to the results section, the main elements to be described are recruitment, participant flow, baseline data, numbers analyzed, the estimated effect size and its precision (such as 95 % confidence interval) for each primary and secondary outcome, results of ancillary analysis and finally harms. Thus, the results section accounts for the flow of the trial as well as the participant outcomes. While symptom data are certainly going to be the core of any N-of-1 trial report, that data cannot be interpreted in isolation.

The results section should begin with a clear account of the number and sequence of periods completed and any changes from the original plan and the reasons for those changes. If the report describes a series of N-of-1 trials, the number of participants who were enrolled, assigned to interventions and analyzed for the primary outcome should be described. Any losses or exclusion of participants after treat-

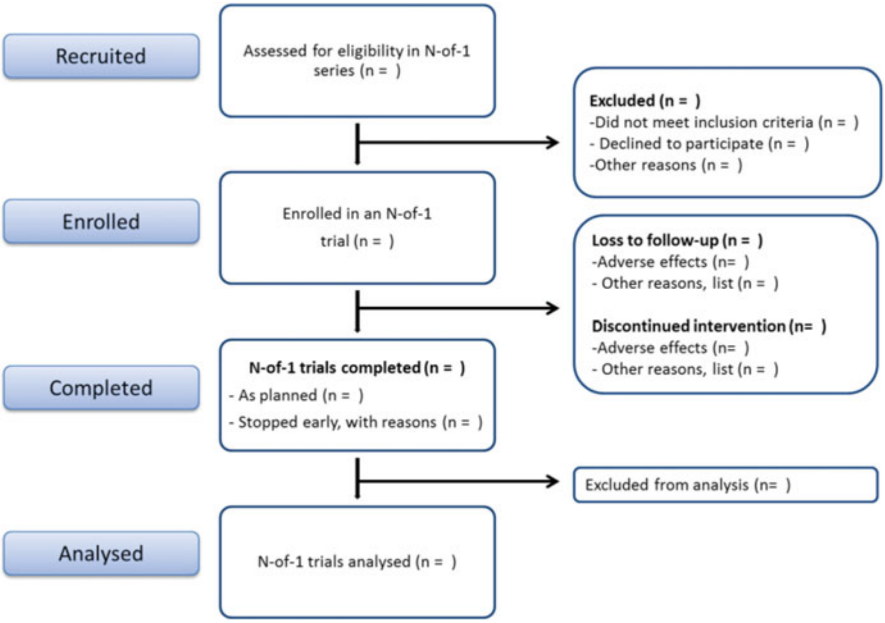

Recruited
Assessed for eligibility in N-of-1 series (n =)
Excluded (n =)
-Did not meet inclusion criteria (n =)
- Declined to participate (n =)
-Other reasons (n =)
Enrolled
Enrolled in an N-of-1 trial (n =)
Loss to follow-up (n =)
-Adverse effects (n=)
- Other reasons, list (n =)
Discontinued intervention (n=)
-Adverse effects (n=)
- Other reasons, list (n =)
Completed
N-of-1 trials completed (n =)
- As planned (n =)
- Stopped early, with reasons (n =)
Excluded from analysis (n=)
Analysed
N-of-1 trials analysed (n =)

Fig. 14.1 Example of participant flow diagram from an N-of-1 series

ment assignment should be accounted for. Whether or not any periods were stopped early and whether or not the trial was stopped early should be reported and any early stopping should be explained.

A diagram showing the flow of participants through the trial is strongly recommended. This will be similar to the flow diagram recommended by the CONSORT statement and an example from an N-of-1 series is presented in the CENT E&E and reproduced here (Fig. 14.1) (Shamseer et al. 2015).

Following the description of the trial flow, a table showing baseline demographic and clinical characteristics for trial participants is needed.

This brings us to the heart of the results section: reporting the results for the outcomes of interest. This includes stating the estimated effect size (i.e. the magnitude of change in outcome for one treatment compared to another) and its precision (e.g. 95 % confidence interval) for each primary and secondary outcome. For binary outcomes, presentation of both absolute and relative effect sizes is recommended. In addition, for a series of N-of-1 trials where a quantitative synthesis was performed, group estimates of effect and precision for each primary and secondary outcome should be stated.

Since N-of-1 trials consist of repeated periods, and sometimes multiple outcome measurements within periods, authors may first summarize and present data for each treatment before estimating effect size.

In addition to a table or text containing this information, authors may also present it in the form of a simple graph, plotting each outcome over time, distinguishing

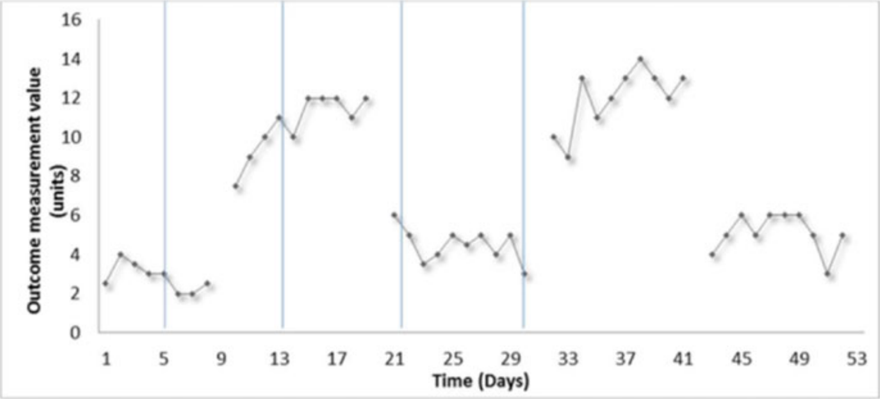


Fig. 14.2 Example of a trial pictorial showing outcome over time for an N-of-1 trial

treatment and comparator. One possible trial pictorial an individual trial is presented in the CENT E&E and reproduced here (Fig. 14.2) (Shamseer et al. 2015). Outcome data are plotted on the y-axis with unit of time (days, weeks, etc.) along the x-axis with vertical lines or some other distinction (e.g. shading) separating treatment periods and pairs or treatment blocks. All trial outcomes may be presented together in one graph (using distinguishing plot points) or in individual graphs for each outcome.

For a series, authors may wish to plot outcome data for multiple participants on a single graph or provide individual trial pictorials, in an appendix if necessary. Small multiples can be a very effective way to present such data (Tufte 1990).

It is difficult to estimate true differences between treatments by visual assessment alone or with just raw or summary data for each period/treatment. As such, we strongly encourage authors to also report the effect size for each primary and secondary outcome and each estimate should be accompanied by a measure of precision (i.e. 95 % confidence interval).

Authors should be explicit about what comparisons were made - between data for each period within a pair or treatment block, between data for each treatment (i.e. combined periods within a block) within a block, or between each treatment overall (i.e. combined periods and blocks). A diagram is suggested. Authors should also be explicit about assumptions and adjustments made to account for period or carry over effect. For instance, authors should report whether a carry over effect was explored and how it was accounted for in the analysis.

In a series of N-of-1 trials, effect estimates may be calculated for each individual trial or individual data may be pooled into a group estimate (akin to a meta-analysis). If the former, authors may wish to present analyses for all participants in one figure.

Authors are encouraged to provide all raw data for an individual or series of trials, possibly in an appendix, consistent with the current movement in parallel-group trials (Donegan et al. 2013). Making individual patient data available may facilitate future inclusion of N-of-1 trials in meta-analyses (i.e., using individual patient data) (Riley et al. 2010). Where the nature of the data is such that presenting all data points is prohibitive, authors are encouraged to consider data deposition, discussed later in this chapter.

Results should be reported for all planned primary and secondary end points, and for all participants, not just for analyses that were statistically significant or interesting. Selective reporting of outcomes within population-based RCTs is a widespread and serious problem (Chan et al. 2004) although it is, in many cases, unintentional (Smyth et al. 2011). Trial registration has made selective reporting easier to detect, although it has not eliminated it (Huić et al. 2011).

Results of any other analyses performed, including assessment of carryover effects, period effects and intra-subject correlation, should be reported. Where a series of N-of-1 trials is reported, results of any sub-group analysis that was done should be reported.

All harms or unintended effects for each intervention should be described. If no harms were observed, this should be stated. Without such a statement the reader cannot determine if the treatment was free of unintended effects or if the authors have simply not reported on this important aspect of trial outcomes. Specific guidance for reporting harms associated with trials is available in CONSORT for harms (Ioannidis et al. 2004).

Writing the Discussion

The discussion section, like the introduction, follows the same format for N-of-1 trials as for parallel group trials. Authors should discuss limitations of the study, generalizability of the findings and interpretation of the findings (balancing benefits and harms) taking into consideration other relevant evidence.

Supplemental Information

Finally, some supplemental information is recommended for full transparency; the registration number and name of the trial registry used and where the full protocol can be accessed. Sources of funding and other support should be described. The level of involvement by a funder and their influence on the design, conduct, analysis and reporting of a trial should also be described. If the funder had no such involvement, the authors should state that. Any other sources of support, such as supply of materials and any role of these in the analysis of data and writing of the manuscript should be reported (Moher et al. 2010b).

Trial Registration, Protocol and Data Deposit: Pieces of the Puzzle

Greatest transparency is achieved when a trial is prospectively registered in a public registry and the protocol made available, sometimes through publication. An example is the protocol for an aggregated series of N-of-1 trials of pilocarpine drops to help dry mouth in palliative care patients (Nikles et al. 2013). This protocol is published in an open access journal and the article cites the Australia and New Zealand Clinical Trial Registry Number assigned when the trial was registered. This enables interested readers to compare the outcomes that were to be assessed and the analyses that were planned with those reported for the completed trial.

The Declaration of Helsinki states, "Every research study involving human subjects must be registered in a publicly accessible database before recruitment of the first subject" (Paragraph 35, World Medical Association General Assembly 2013). Further, "the design and performance of each research study involving human subjects must be clearly described and justified in a research protocol" (Paragraph 22). N-of-1 trials can be registered in registers such as ClinicalTrials.gov or in other national trials registries. Most registries have no costs associated with their use and are intended to have at least summary results deposited on completion of the trial.

The objective of trial registries is to make known what trials have been conducted and what they measured and to promote public availability of trial results without either entire studies or certain outcomes being kept from view. However, registration alone is not enough to prevent selective outcome reporting or analysis reporting in registered trials (Dwan et al. 2011, 2013), although it makes it easier to detect (Norris et al. 2014).

The AllTrials Manifesto calls for registration, summary reporting of primary and secondary outcomes within a year of trial completion and full reports made public – redacting only narrative details of adverse events and any other patient identifying material (AllTrials Campaign 2013). AllTrials is an initiative of Sense About Science, Bad Science, BMJ, James Lind Initiative, the Centre for Evidence-based Medicine, PLOS, the Cochrane Collaboration and in the US Dartmouth's Geisel School of Medicine and the Dartmouth Institute for Health Policy & Clinical Practice. Its objective is to ensure that all trials are registered and published and that the reports that are made available to regulators are also made available to other scientists (Goldacre 2014). Such availability opens science to greater scrutiny and provides a richer research base, expanding the potential for collaboration and ensuring that the same study is not unknowingly conducted twice (House of Commons Science and Technology Committee 2013). Although recognizing the opportunities afforded by such data sharing, the AllTrials Manifesto stops short of calling for individual patient data to be made publicly available (AllTrials Campaign 2013).

AllTrials has this to say of trials that have not been reported, or in some cases never registered; "Information on what was done and what was found in these trials could be lost forever to doctors and researchers, leading to bad treatment decisions, missed opportunities for good medicine, and trials being repeated."

Funders or universities may oblige investigators to deposit their data (Science.gc.ca 2011) and some journals may require authors to share their data on request (Hrynaszkiewicz et al. 2010). There are several options for doing this. Many universities have institutional repositories, which can archive dissertations, publications and, increasingly, data sets. Public repositories exist, often focused in particular disciplines (Science.gc.ca 2011). These include Global Biodiversity Information Facility (Edwards 2004), GenBank (Anon 2013), and Dryad, a repository specializing in data sets supporting published, peer reviewed articles in bioscience (Hrynaszkiewicz and Cockerill 2012). Closest at hand for authors, journals will often have the facility to make supplemental information available online. When depositing data with a journal, the author should consider whether assignment of copyright for the article would also apply to the accompanying material. It is interesting to note that, generally, a fact cannot be copyrighted, and so data sets, as collections of facts, do not warrant the same inherent intellectual property protection that a creative work such as a journal article would (Hrynaszkiewicz and Cockerill 2012). A creative commons copyright option can allow for re-use of the data. The CC BY license allows unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited (Creative Commons 2014). A CC0 license places the data set entirely in the public domain (Creative Commons 2014), and has been recommended for biomedical datasets associated with journal articles (Hrynaszkiewicz and Cockerill 2012).

A benefit of data deposit with the journal is that there will be a direct link from the article to the associated data and that the data will become available when the article is published, but not before. However, repositories such as Dryad will provide a DOI – digital object identifier – that can be published in the article, effectively providing such a link (American Psychological Association 2012), as well as placing an embargo on the dataset until the publication of the associated article. It should be noted that there is widespread agreement from journal editors that such deposit does not constitute prior publication (Krleza-Jerić and Lemmens 2009).

Data files should be in a generic format such as a comma delimitated CSV text file format rather than the proprietary format of a statistical analysis package. As well, documentation of the variables and their coding should be made available (Hrynaszkiewicz and Cockerill 2012).

In making datasets publically available, confidentiality of the study participants must be assured and some data preparation will be required to provide or confirm de-identification. De-identification must follow all regulatory requirements, such as the Health Insurance Portability and Accountability Act (HIPAA) in the United States, and must consider both direct and direct identifiers. Indirect identifiers are those that could potentially identify a participant when used in combination (Hrynaszkiewicz et al. 2010).

The need for de-identification may present particular challenges for data deposit for N-of-1 trials. For situations where research participants have highly stigmatized conditions, such as human immunodeficiency virus infection, or have a rare condition which is itself identifying, the privacy risk may be too great to warrant data deposit (El Emam et al. 2012) although some do advocate for data sharing as necessary to advance the treatment of rare diseases (Pleticha 2014). In datasets that are to be publicly available, rather than available to those with certain credentials, and which contain data for small numbers of participants, the risk of re-identification is considered high (El Emam 2008). The privacy afforded by removing identifying information from the dataset will be entirely undone if the information supplied in the manuscript concerning location of the research or characteristics of the participants can be easily collated with dataset elements to re-identify those participants. When there is any doubt about the protection of the research participant, authors should discuss the data release with their institutional review board (Hrynaszkiewicz et al. 2010).

Clearly data deposit must be planned in advance. Research participants must be informed in the consent process and data management planned to ensure it can be de-identified and fully documented without substantial additional effort.

Conclusion

A full report of clinical data to professional audiences will include trial registration, a publicly available protocol, journal publication of methods and results without bias toward positive findings as well as public deposit of the anonymized clinical data. The journal article is the core of this reporting as it is often the only product of a research study that is available to the public, with the other elements ensuring full transparency and data reusability for a research study overall. CENT provides a structured format to ensure that the main journal report is sufficiently detailed that it can be critically appraised and replicated.

References

AllTrials Campaign (2013) All trials Manifesto. http://www.alltrials.net/all–trials/. Available at: http://www.alltrials.net/all-trials/. Accessed 8 May 2014

American Psychological Association (2012) Electronic sources and locator information. In: Publication manual of the American Psychological Association. American Psychological Association, Washington, pp 187–188. Available at: http://www.apastyle.org/learn/faqs/what-is-doi.aspx

Anon (2013) GenBank overview. National Center for Biotechnology Information, U.S. National Library of Medicine. Available at: http://www.ncbi.nlm.nih.gov/genbank/. Accessed 12 May 2014

Begg C et al (1996) Improving the quality of reporting of randomized controlled trials. The CONSORT statement. JAMA 276(8):637–639

Boutron I et al (2008a) Extending the CONSORT statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. Ann Intern Med 148(4):295–309

Boutron I et al (2008b) Methods and processes of the CONSORT Group: example of an extension for trials assessing nonpharmacologic treatments. Ann Intern Med 148(4):W60–W66

Calvert M et al (2013) Reporting of patient-reported outcomes in randomized trials: the CONSORT PRO extension. JAMA 309(8):814–822. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23443445. Accessed 10 July 2014

Chan A-W et al (2004) Empirical evidence for selective reporting of outcomes in randomized trials: comparison of protocols to published articles. JAMA 291(20):2457–2465. Available at: http://jama.jamanetwork.com/article.aspx?articleid=198809. Accessed 6 May 2014

Creative Commons (2014) About the licenses – Creative Commons. Available at: https://creativecommons.org/licenses/. Accessed 12 May 2014

Creative Commons (2014) About CC0 public domain dedication – "no rights reserved." Available at: http://creativecommons.org/about/cc0. Accessed 12 May 2014

Donegan S et al (2013) Combining individual patient data and aggregate data in mixed treatment comparison meta-analysis: individual patient data may be beneficial if only for a subset of trials. Stat Med 32(6):914–930

Dwan K, Altman DG, Cresswell L, Blundell M, Gamble CL, Williamson PR (2011) Comparison of protocols and registry entries to published reports for randomised controlled trials. Cochrane Database Syst Rev 1, MR000031. doi:10.1002/14651858.MR000031.pub2.

Dwan K et al (2013) Systematic review of the empirical evidence of study publication bias and outcome reporting bias – an updated review. PloS One 8(7):e66844. Available at: http://www.plosone.org/article/info:doi/10.1371/journal.pone.0066844#pone-0066844-g022. Accessed 29 Apr 2014

Edwards JL (2004) Research and societal benefits of the global biodiversity information facility. BioScience 54(6):485–486

El Emam K (2008) Heuristics for de-identifying health data. IEEE Secur Priv 6(4):58–61

El Emam K et al (2012) De-identification methods for open health data: the case of the Heritage Health Prize claims dataset. J Med Internet Res 14(1):e33. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3374547&tool=pmcentrez&rendertype=abstract. Accessed 1 May 2014

Gage NA, Lewis TJ (2013) Analysis of effect for single-case design research. J Appl Sport Psychol 25(1):46–60. Available at: http://dx.doi.org/10.1080/10413200.2012.660673. Accessed 29 Apr 2014

Gagnier JJ et al (2006) Reporting randomized, controlled trials of herbal interventions: an elaborated CONSORT statement. Ann Intern Med 144(5):364–367. Available at: http://www.sciencedirect.com/science/article/pii/S0895435606002502. Accessed 5 May 2014

Goldacre B (2014) AllTrials. Available at: http://www.alltrials.net. Accessed 5 May 2014

Guyatt GH et al (2000) Users' guides to the medical literature: XXV. Evidence-based medicine: principles for applying the users' guides to patient care. JAMA 284(10):1290–1296. Available at: http://jama.ama-assn.org/cgi/doi/10.1001/jama.284.10.1290

Haynes RB et al (1990) More informative abstracts revisited. Ann Intern Med 113(1):69–76

Hoffmann TC et al (2014) Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ (Clinical research ed.) 348(mar07_3):g1687. Available at: http://www.bmj.com/highwire/filestream/689456/field_highwire_article_pdf/0/bmj.g1687. Accessed 29 Apr 2014

Hopewell S et al (2008) CONSORT for reporting randomized controlled trials in journal and conference abstracts: explanation and elaboration. PLoS Med 5(1):e20. Available at: http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.0050020#pmed-0050020-t001

Horner RH et al (2012) Considerations for the systematic analysis and use of single-case research. Educ Treat Child 35(2):269–290. Available at: http://muse.jhu.edu/journals/education_and_treatment_of_children/v035/35.2.horner.html. Accessed 5 May 2014

House of Commons Science and Technology Committee (2013) Clinical trials. The Stationary Office Limited, London

Hrynaszkiewicz I, Cockerill MJ (2012) Open by default: a proposed copyright license and waiver agreement for open access research and data in peer-reviewed journals. BMC Res Notes 5(1):494. Available at: http://www.biomedcentral.com/1756-0500/5/494. Accessed 1 May 2014

Hrynaszkiewicz I et al (2010) Preparing raw clinical data for publication: guidance for journal editors, authors, and peer reviewers. Trials 11(1):9. Available at: http://www.trialsjournal.com/content/11/1/9. Accessed 29 Apr 2014

Huić M, Marušić M, Marušić A (2011) Completeness and changes in registered data and reporting bias of randomized controlled trials in ICMJE journals after trial registration policy. N. Siegfried, ed. PloS One 6(9):e25258. Available at: http://dx.plos.org/10.1371/journal.pone.0025258. Accessed 1 May 2014

Huth EJ (1987) Structured abstracts for papers reporting clinical trials. Ann Intern Med 106(4):626. Available at: http://annals.org/article.aspx?articleid=701794. Accessed 9 May 2014

International Committee of Medical Journal Editors (1997) Uniform requirements for manuscripts submitted to biomedical journals. N Engl J Med 336(4):309–315. Available at: http://linkinghub.elsevier.com/retrieve/pii/S0168827897803066. Accessed 5 May 2014

Ioannidis JPA et al (2004) Better reporting of harms in randomized trials: an extension of the CONSORT statement. Ann Intern Med 141(10):781–788

Irwig L, Glasziou P, March L (1995) Ethics of N-of-1 trials. Lancet 345(8948):469. Available at: http://www.ncbi.nlm.nih.gov/pubmed/7861872. Accessed 5 May 2014

JAMA (2014) Instructions for authors. Available at: https://jama.jamanetwork.com/public/instructionsForAuthors.aspx#CONSORTFlowDiagramandChecklist. Accessed 6 May 2014

Jones J, Hunter D (1995) Consensus methods for medical and health services research. BMJ 311(7001):376–380. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2550437&tool=pmcentrez&rendertype=abstract. Accessed 2 May 2014

Krleza-Jerić K, Lemmens T (2009) 7th revision of the declaration of Helsinki: good news for the transparency of clinical trials. Croat Med J 50(2):105–110. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC2681053/. Accessed 1 May 2014

Lefebvre C et al (2013) Methodological developments in searching for studies for systematic reviews: past, present and future? Syst Rev 2(1):78. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24066664. Accessed 17 Oct 2013

MacPherson H et al (2010) Revised STandards for Reporting Interventions in Clinical Trials of Acupuncture (STRICTA): extending the CONSORT statement. J Evid Based Med 3(3):140–155

Mahon JL, Feagan BG, Laupacis A (1995) Ethics of N-of-1 trials. Lancet 345(8955):989. Available at: http://www.sciencedirect.com/science/article/pii/S0140673695907386. Accessed 5 May 2014

Moher D (2009) Guidelines for reporting health care research: advancing the clarity and transparency of scientific reporting. Can J Anaesth 56(2):96–101

Moher D, Schulz KF, Altman DG (2001) The CONSORT statement: revised recommendations for improving the quality of reports of parallel group randomized trials. BMC Med Res Methodol 1:2. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=32201&tool=pmcentrez&rendertype=abstract. Accessed 10 Jul 2014

Moher D et al (2009) Preferred reporting items for systematic reviews and meta-analyses: the PRISMA statement. PLoS Med 6(7):e1000097. Available at: http://www.plosmedicine.org/article/info%3Adoi%2F10.1371%2Fjournal.pmed.1000097

Moher D et al (2010a) CONSORT 2010 explanation and elaboration: updated guidelines for reporting parallel group randomised trials. BMJ 340:c869–c869. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.c869. Accessed 6 Jul 2010

Moher D et al (2010b) Guidance for developers of health research reporting guidelines. PLoS Med 7(2):e1000217. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2821895&tool=pmcentrez&rendertype=abstract

Moher D et al (2011) Describing reporting guidelines for health research: a systematic review. J Clin Epidemiol 64(7):718–742

Morris C (2008) The EQUATOR network: promoting the transparent and accurate reporting of research. Dev Med Child Neurol 50(10):723

Nakayama T et al (2005) Adoption of structured abstracts by general medical journals and format for a structured abstract. J Med Libr Assoc 93(2):237–242. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=1082941&tool=pmcentrez&rendertype=abstract. Accessed 29 Apr 2014

Nikles J et al (2013) Do pilocarpine drops help dry mouth in palliative care patients: a protocol for an aggregated series of n-of-1 trials. BMC Palliat Care 12(1):39. Available at: http://www.biomedcentral.com/1472-684X/12/39. Accessed 29 Apr 2014

Norris SL et al (2014) Clinical trial registries are of minimal use for identifying selective outcome and analysis reporting. Res Synth Met. Available at: http://doi.wiley.com/10.1002/jrsm.1113. Accessed 19 Mar 2014

Pleticha R (2014) Guest post: people with rare diseases need results from all trials. All Trials News. Available at: http://www.alltrials.net/2014/guest-post-people-with-rare-diseases-need-results-from-all-trial/. Accessed 12 May 2014

Price JD, Grimley Evans J (2002) N-of-1 randomized controlled trials (‘N-of-1 trials’): singularly useful in geriatric medicine. Age Ageing 31(4):227–232

Punja S et al (2014) Ethical framework for N-of-1 trials: clinical care, quality improvement, or human subjects research? In: Kravitz R, Duan N, De. M. C. N.-1G. Panel (eds) Design and implementation of N-of-1 trials: a user’s guide. Agency for Healthcare Research and Quality, Rockville, pp 13–22

Riley RD, Lambert PC, Abo-Zaid G (2010) Meta-analysis of individual participant data: rationale, conduct, and reporting. BMJ 340:c221. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.c221. Accessed 6 May 2014

Ripple AM et al (2011) A retrospective cohort study of structured abstracts in MEDLINE, 1992–2006. J Med Libr Assoc 99(2):160–163. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3066587&tool=pmcentrez&rendertype=abstract. Accessed 29 Apr 2014

Schardt C et al (2007) Utilization of the PICO framework to improve searching PubMed for clinical questions. BMC 7(1):16. Available at: http://www.biomedcentral.com/1472-6947/7/16. Accessed 7 May 2014

Schulz KF, Altman DG, Moher D (2010) CONSORT 2010 Statement: updated guidelines for reporting parallel group randomised trials. BMJ 340:c332–c332. Available at: http://www.bmj.com/cgi/doi/10.1136/bmj.c332. Accessed 13 Apr 2011

Science.gc.ca (2011) Open access: research data. Available at: http://www.science.gc.ca/default.asp?lang=en&n=2BBD98C5-1. Accessed 12 May 2014

Shamseer L, Sampson M, Bukutu C, Schmid CH, Nikles J, Tate R, Johnston BC, Zucker D, Shadish WR, Kravitz R, Guyatt G, Altman DG, Moher D, Vohra S, CENT group (2015) CONSORT extension for reporting N-of-1 trials (CENT) 2015: explanation and elaboration. BMJ 350:h1793. doi:10.1136/bmj.h1793

Sharma S, Harrison JE (2006) Structured abstracts: do they improve the quality of information in abstracts? Am J Orthod Dentofacial Orthop 130(4):523–530. Available at: http://www.sciencedirect.com/science/article/pii/S0889540606008936. Accessed 29 Apr 2014

Smyth RMD et al (2011) Frequency and reasons for outcome reporting bias in clinical trials: interviews with trialists. BMJ 342:c7153. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3016816&tool=pmcentrez&rendertype=abstract. Accessed 6 May 2014

Sollaci LBMGP (2004) The introduction, methods, results, and discussion (IMRAD) structure: a fifty-year survey. J Med Libr Assoc 92(3):364–367. Available at: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC442179/. Accessed 29 Apr 2014

Stevenson HA, Harrison JE (2009) Structured abstracts: do they improve citation retrieval from dental journals? J Orthod 36(1):52–60; discussion 15–6

Tufte E (1990) Envisioning information. Graphics Press, Cheshire

Turner L, Shamseer L, Altman DG, Weeks L et al (2012a) Consolidated standards of reporting trials (CONSORT) and the completeness of reporting of randomised controlled trials (RCTs) published in medical journals. Cochrane Database Syst Rev 11(11):MR000030. Available at: http://www.ncbi.nlm.nih.gov/pubmed/23152285. Accessed 8 May 2014

Turner L, Shamseer L, Altman DG, Schulz KF et al (2012b) Does use of the CONSORT Statement impact the completeness of reporting of randomised controlled trials published in medical journals? A Cochrane review. Syst Rev 1:60. Available at: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3564748&tool=pmcentrez&rendertype=abstract

Vohra S, Shamseer L, Sampson M, Bukutu C, Schmid CH, Tate R, Nikles J, Zucker DR, Kravitz R, Guyatt G, Altman DG, Moher D, CENT group (2015) CONSORT extension for reporting N-of-1 trials (CENT) 2015 statement. BMJ 350:h1738. doi:10.1136/bmj.h1738

Wen J et al (2008) The reporting quality of meta-analyses improves: a random sampling study. J Clin Epidemiol 61(8):770–775. Available at: http://www.ncbi.nlm.nih.gov/pubmed/18411041. Accessed 27 Nov 2010

Wilczynski NL et al (1995) Preliminary assessment of the effect of more informative (structured) abstracts on citation retrieval from MEDLINE. MEDINFO 8(Pt 2):1457–1461

World Medical Association General Assembly (2013) Declaration of Helsinki: ethical principles for medical research involving human subjects. JAMA 310(20):2191–2194. Available at: http://www.ncbi.nlm.nih.gov/pubmed/24141714

Chapter 15
Single Patient Open Trials (SPOTs)

Jane Smith, Michael Yelland, and Christopher Del Mar

Abstract Single patient open trials (SPOTs) are nearly identical to standard trials of treatment. The added essential ingredient is a set of symptoms (commonly arrived at by negotiation between clinician and patient) to monitor (the *outcome measures*). This means they lie somewhere in between formal N-of-1 trials and totally informal trials of treatment in terms of rigour. SPOTs are accordingly less demanding to arrange (for both the patient and clinician) than N–of-1 trials, but they require considerably more effort and commitment than casual trials of treatment. This chapter defines and describes the rationale for SPOTs, discusses when and why they could be used, as well as their limitations, and describes outcome measures and analysis. As well as describing the use of SPOTs in clinical contexts, it covers the extra considerations required when using SPOTs in research. Several examples of the practical application of SPOTs are given, some with the resulting data. It is anticipated that the examples may be adapted to enable other clinicians and their patients to perform their own SPOTs to validate other medical interventions in the context of the individual.

Keywords Ethics • Experimental • General practitioners • Informed consent • Outcome assessment (treatment outcome) • Patient centred outcome research • Placebo effect • Protocol • Research design • Single patient open trials (SPOTs) • Trials of treatment

J. Smith MBBS, FRACGP. Grad Dip FM, MHS, FAICD (✉) •
C. Del Mar, MD, Bsc, MBBS,FRACGP, FAFPHM
Bond University, Gold Coast, QLD, Australia
e-mail: jsmith@bond.edu.au; cdelmar@bond.edu.au

M. Yelland, MBBS, Ph.D., FRACGP, FAFMM. Grad Dip Musculoskeletal Med
School of Medicine, Griffith University and Menzies Health Institute,
Gold Coast, QLD, Australia
e-mail: m.yelland@griffith.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_15

Where the Notion Came From

We dedicate this chapter to Professor Charles Bridges-Webb (Brown 2010). He was interested in N-of-1 trials as a means to increase the research exposure of general practitioners (GPs), and recognised the difficulty of creating placebo for them. He suggested we run the trials open rather than blinded: we could still document and quantify symptoms with time, comparing alternative treatment periods (e.g. “Drug a” vs. “Drug b”), or comparing treatment with “Drug a” to nothing.

The Current Clinical Status Quo: ‘Trial of Treatment’

GPs are drawn to pragmatic approaches. One of these is the therapeutic trial (“trial of treatment”), a well-established practical tool for deciding if a treatment is going to be useful: the patient is asked to ‘try’ the treatment, and report back if the symptoms are helped or not. This enables a long term treatment plan to be formulated. This approach has three weaknesses.

Firstly both the placebo effect and the regression-to-mean effects are likely to play a large role. The placebo effect is well known. Patients are likely to *expect* that their symptoms will feel better if offered a new, possibly expensive, treatment presented with hope and perhaps a list of the dangers of this new treatment (perhaps the reason why it hadn’t been contemplated until now). The expectations may well become experiences (Howick et al. 2013).

Secondly regression-to-the-mean, which is less well understood as a statistical phenomenon (Bland and Altman 1994), has important clinical effects. In this setting, because patients tend to seek help when fluctuating symptoms are at their worst, then just the passage of time will result in those symptoms settling. Symptoms that increase and decrease in severity with time will tend to ‘regress’ towards the average severity, after re-measuring any outlier measurement. This means that any new treatment adopted can too easily be assumed to the cause of any improvement in a *post hoc, ergo propter hoc* fallacy.

Finally, the vagueness of the outcome of patients’ symptoms means that any placebo and regression-to-the-mean effects may become amplified. One of the strengths of randomised trials (including N-of-1 trials) is that great care is taken to identify *outcome measures* – in this case, symptoms to follow with time. Otherwise our all-too-human characteristics of over-estimating causal effects mean that we tend to focus on symptom components that improve, and neglect those that do not (or even get worse) if they do not conform to the expected benefit pattern.

Box 15.1: Essential Components of SPOTs

1. They have the purpose of choosing the best of two (or possibly more) management options for a single patient;
2. Each of these options is clearly defined, and reproducible;
3. The treatment options may be alternated several times depending on the patient preferences to allow comparison of the symptoms experienced between each time period;
4. There is a 'washout' period between each treatment period to reduce any lagging treatment effect contaminating the next treatment time period.
5. Patients record their symptoms as objectively as possible to allow comparisons.
6. At the end of the SPOT, clinician and patient make a decision about the benefits and adverse effects and decide which treatment to continue with.

What Is a Single Patient Open Trial (SPOT)?

Single patient open trials (SPOTs) are nearly identical to trials of treatment. The added essential ingredient is a (commonly arrived at) set of symptoms to monitor (the *outcome measure*). This means they lie somewhere in between formal N-of-1 trials and totally informal trials of treatment in terms of rigour. SPOTs are accordingly less demanding to arrange (for both the patient and clinician) than N of 1 trials, but they require considerably more effort and commitment than casual trials of treatment.

SPOTs have several essential components (see Box 15.1).

In summary, SPOTs provide a framework and methodology for evaluating the patient response in single patients in practice using patient centred outcomes to assess the extent of benefit of any given treatment in an individual.

Why Do We Need SPOTs?

SPOTs can help in situations in which there is uncertainty about two or more treatment options – in exactly the same way that N-of-1 trials can help.

Why should there be uncertainty? One could argue that we have the whole edifice of evidence-based medicine (EBM) to help direct us to the most effective treatment. This is true, and for many decisions, we can go to trials for the best available evidence to help us choose the most effective, with the least cost (in terms of adverse effects, as well as financial cost).

However there are some situations in which *individual variation* plays an important role. Individual patients vary in their response to some treatments: what is effective for one patient may not be effective for another. This is presumably because of different genetically determined biochemical differences in metabolism. Usually this is not elucidated, and we have to resort to this form of empirical testing. One day, we may be able to undertake specific tests in individual patients to decide whether they will respond to a certain drug or not.

A well-known example of such a clinical situation is that of the disease of osteoarthritis of the joints, and the possible alternatives of paracetamol (acetaminophen) or non-steroidal anti-inflammatory drugs (NSAIDs). The conundrum facing the prescribing clinician is which of these two different classes to use. Paracetamol is generally regarded as much safer than the NSAIDs, which cause gastro-intestinal bleeding, fluid retention and a higher risk of cardiovascular events such as myocardial infarction. This means that most primary care doctors will start patients on paracetamol first. If this does not control the pain and stiffness, then consider stepping up to one of the NSAIDs, perhaps employing a trial of treatment to do so. However we know that less than half of patients will respond to NSAIDs (March et al. 1994). The rest will be exposed to the extra risk with no benefit. How can we decide which group the patient in front of us belongs to?

The best method of course would be an N-of-1 trial (March et al. 1994). However, this is beyond most clinicians (especially in primary care, where this decision is faced most commonly). A less rigorous method is a trial of therapy, for the reasons outlined above.

SPOTs then become a second-best (but much more accessible) method.

When Should a SPOT Be Considered?

SPOTs can help make clinical decisions when patient and clinician are uncertain about the best of two or more different treatment options.

Firstly there needs to be a willingness of both clinician and patient to recognise the uncertainty, and join together to answer the question. If either have a strong view about the efficacy of one of the choices (that is, therapeutic equivalence does not exist), then this is not a suitable option. Both must be prepared to expend the extra effort in collecting the data and attempting to interpret it later, and the patient (particularly) to endure periods of time using what might turn out to be the less effective treatment option.

Secondly the disease state must be stable, and unlikely to spontaneously resolve. This means that SPOTs are limited to chronic conditions with persistent symptoms (or possibly other markers of disease, such as blood pressure, or glycolated haemoglobin [HbA1c]). However it does not matter if these symptoms fluctuate (it just means the SPOT might have to take place over a longer time).

Table 15.1 Comparison of disease centred versus patient-centred management of SPOT processes

SPOT processes	Disease-centred/bio-medical questions	Patient-centred questions
Implementation	Will the patient adhere/comply?	Does the patient like the treatment enough to use it?
Efficacy	Symptom and or disease control	Does it address the patient's concerns?
Drug response?	Is this a placebo effect?	Does the patient feel better?
Safety	Are there adverse effects?	Are (any) unwanted effects or safety risks enough to outweigh (any) benefits?

SPOTs and Patient Centered Outcomes

SPOTs are by their nature a way of practicing *patient-centred* medicine. Patient-centeredness is a process of focusing the centre of the consultation on the patient as a person rather than the collection of disease within. Consulting in a patient centred way leads to the identification of patient centred outcomes. Its contrast is with the 'disease-centred', or the 'bio-medical approach', all too often a throwback to earlier medical school teaching, Table 15.1. Its fundamental precept is that optimal care can only be delivered by understanding the fears, concerns and expectations of the patient (Stewart 1995).

The SPOT process requires a doctor-patient discussion that explores the patient's perspective on their illness. Some of its key components can be listed (Stewart 2005):

- Exploring the problems presented from both the patient's, as well as the disease, viewpoint
- Understanding the whole person
- Finding common ground
- Improving the patient-doctor relationship
- Being realistic.

It is only possible to undertake a SPOT if clinician and patient can communicate effectively: they need to choose and agree on a patient centred outcome to use. This process is called *shared decision-making* (Barry and Edgman-Levitan 2012; Yelland and Schluter 2006) – the extension of patient centeredness to management and treatment. The importance of patient centeredness and shared decision-making are teased out below.

Conventional Therapeutic Approach Compared to SPOT

Traditional thinking about therapy focuses on whether patients adhere to or comply with treatments. This is a top-down, paternalistic approach. In contrast, SPOTs elicit the patient's preferences as part of the process. This means that patients may be more likely to continue the treatment if they are instrumental in gathering data for its evaluation on themselves, and rate it better than the alternative.

However it should be pointed out that there is little empirical evidence for this effect. Ideally this research question will be addressed in the future.

Limitations of SPOTs

There are problems with SPOTs. The most important, discussed above, is that the patient is not blind to which treatment they are using at each time period. It is too easy to attribute to a treatment incorrectly, effects that are in fact coming from either a placebo response or regression to the mean.

Although SPOTs may be more patient-centred than N-of-1 trials, the level of evidence they provide on the effectiveness of treatments is lower due to the methodological differences highlighted in Table 15.2.

Some of these shortcomings can be partly mitigated by a focus on making the outcome as robust as possible. The vague outcomes (e.g. “feeling better”; “sleeping better”) that are tolerated in informal trials of treatment, are often too subjective to be effective. The identification of an objective-enough outcome to measure, that still has meaning to the patient, may be challenging, and may require careful probing during the consultation (perhaps extending over more than one).

Conditions and Treatments That Are Suitable for SPOTs

SPOTs are not suitable for all patients and all conditions – the condition must be stable and long standing enough to be able to measure the effects of the interventions over a reasonable period with minimal variation attributable to natural history. The patient needs to have sufficient interest in their condition and its optimum

Table 15.2 Comparison of SPOT and N-of-1 methodology

	N-of-1	SPOT
Research method	Randomisation of treatment periods for treatment, placebo control, and/or comparator Blinding of patient and doctor to treatments Precision from use of three or more crossovers	No randomisation No placebo control No blinding One or more crossovers
Timelines	Pre-specified protocol controlled treatment crossovers	Patient controlled: doctor and patient agreed
Outcome measures	Validated measures of effect	Patient and clinician agree on outcomes to be measured
Analysis	Positive result determined by evidence on minimum clinically important differences	Positive result defined by doctor and patient, preferably a priori Less certain result

Table 15.3 Examples of conditions and treatments suitable for SPOTs

Condition	Comparison treatments
Acne vulgaris	Topical benzoyl peroxide vs. topical clincamycin
Asthma	Montelukast vs. inhaled corticosteroid
Chronic osteoarthritis of the knee	NSAID vs. paracetamol
Insomnia	Valerian vs. camomile tea
Leg cramps	Oral magnesium orotate vs. oral calcium carbonate
Premenstrual syndrome	Vitex agnus castus vs. sertraline
Tension headache	Oral magnesium vs. no treatment
Vomiting in infancy	Avoiding dairy products or normal diet in breastfeeding mother

treatment to be actively involved in its planning and execution. Systematically recording outcomes over a considerable period is not a trivial impost.

Clearly there must be two or more potentially effective treatments to compare. One option may also be no treatment. Remember that an adverse effect from a treatment may be the focus of the SPOT, and this is just as valid a reason as the beneficial response of a treatment.

Several examples of potential SPOTs are given in Table 15.3.

Outcome Measures

What to Measure

The indicator or outcome needs to be clear, quite specific, and reflect what is most important to the patient. Then this choice must be agreed. The more the outcome is relevant to the patient, the more likely they will be interested enough to document their responses e.g. level of pain, mobility, unwanted fatigue or other drug side effects. Outcome measures may include both subjective (e.g. patient reported outcomes) and objective (e.g. respiratory function tests). The number of measures should be kept to the minimum required to achieve the aims of the SPOT.

How to Measure it

In addition to what outcome to measure, the GP and patient need to agree on a means to measure the outcome, and how to document the response. Symptoms or other measurements can be measured and documented. Examples are: pain (quantified as points on a 0–10 numerical scale); blood glucose (absolute value); and blood pressure self-measurements (absolute value).

Documenting the Outcome

There are many recording techniques for outcome measurement. Examples include: manually entering information in a written or electronic diary format; and by using advanced technology available in smart phone apps to provide both automatic measurements and simultaneous storage on smart phones.

Using Apps as Measuring Tools

There are an abundance of apps for all sorts of things including measurement of exercise, diet, heart rate, sleep and mental health, available free or at a small cost. Their functionality and their usefulness to the SPOT process varies; just how much they have to offer depends on both their appeal to an individual patient and if what they measure fits with the chosen outcome/s to be measured.

This means that if heart rate is an issue for patient and clinician, then an App such as "Instant Heart rate" can be used to measure and record pulse rate at specified times. Alternatively if sleep or mood is the issue, then there are Apps to record these values too. See Table 15.4 for examples of these.

When to Measure, and for How Long

The frequency and duration of monitoring need to be agreed between patient and doctor. Just how often it is practical or realistic to measure, as well as how long for (e.g. twice a day for 2 weeks; or once a day for 1 month) will depend on the patient's interest and attitude. Greater scientific rigour is more likely to be obtained by repeated cycles of treatment versus comparator and/or no treatment exposures. Example 2 (p. 206) refers to 3 cycles of 4 weeks each. It is important to adjust around the predicted length of drug washout periods, and any expected delays in the onset of action of each medication (to avoid misinterpretation of carryover effects).

Table 15.4 Examples of Apps with potential to be used in SPOTs

Feature	Name of app	Measures	Cost
Physical activity	Ride with MapMyRide Run with MapMyRun Walk with MapMyWalk	Distance, (+/−speed) and kilojoules burnt	Free Free Free
Diet	Australian Calorie counter easy diet diary	Energy content of food as kJ	Free
Pulse	Instant Heart Rate	Pulse rate	Free
Mental health	Anxiety iSelfhelp mental health test iMoodJournal	GAD-7 anxiety scale Depression score Mood score (out of 10)	Free Free $0.99
Sleep	Sleep Time	Length and quality of sleep	Free

Symptoms subject to more fluctuation will need more data to off-set the greater uncertainty.

Analysing the Results

Ideally the analysis of the results of SPOTs is both simple and meaningful for patient as well as clinician. This means simple enough to be undertaken without statistical training. Apps used for data collection could be designed to do some of these simple calculations. Descriptive results are usually fairly simple to understand and may just be tallied, e.g. ‘treatment A was preferred over treatment B in two cycles and no preference was expressed in the third cycle’.

In fact the analysis of SPOTs could be as complicated as for N-of-1 trials (see Chap. 12). However in SPOTs more emphasis is placed on the individual patient’s interpretation of the data, and less on what is significant (either at the *statistical* or *minimum clinically important difference* level).

Deciding what is a statistically significant difference between the time periods of an N-of-1 trial or SPOT is often difficult. Some outcomes may be already well-defined by existing clinical guidelines, for example blood pressure targets for diabetics. For some quantitative outcomes, minimum clinically important differences can be found in the clinical research literature, but these are essentially a statistical construct based on a comparison of change scores over time with the patients’ global impression of change for a series of patients (Copay et al. 2007). What patients regard as a worthwhile change in their clinical status can vary widely. For example in chronic low back pain, this can vary from 1 % to 100 % for reductions in pain and in disability (Yelland and Schluter 2006).

Ideally the threshold for a worthwhile difference in response to the treatments in question should be decided in consultation with the patient *before* the SPOT commences to avoid the *post hoc, ergo propter hoc* fallacy. This also fits in well with the patient-centred approach of the SPOT.

For SPOTs with more than one outcome, a more complex process will be needed to make conclusions about differences in effectiveness of treatments. What is the relative value of each outcome medically and to the patient? Should all outcomes be weighted equally or does one take precedence over another? Are the benefits of treatments outweighed by their adverse effects? Ultimately it may be difficult to make a definitive statement about results of a SPOT and it may be more appropriate to make a statement describing the findings for each outcome.

The extent to which the conclusions from SPOTs will influence decisions about future management is unknown at present and should be a priority topic in future research about SPOTs. However it seems a reasonable hypothesis that involving the patient in all stages of the SPOT process from design to data collection to decisions about the results and conclusions, will give them a sense of ownership that may help them to adhere better to management decisions than through the conventional informal trial of treatment.

What About Research?

Up to now, this chapter has been written with clinical care in mind. However SPOTS can be used in research just as N-of-1 trials can.

They can be used in two ways:

1. SPOTs as single 'case reports'. These tend to be hypothesis-generating rather than providing any generalizable information to predict how future patients may be managed;
2. A series of SPOTs. These are more able to be generalised to future patients, as in "what can be expected from patients presenting with equipoise about X treatment for Y disease…".

There are extra considerations for the use of SPOTs as a research tool.

Informed Consent

Of course patients offer informed consent when taking part in a clinical SPOT. However its provision is implicit and assumed. After all, if they decide they do not want to participate or continue, they simply stop.

However in research that is to be published, informed consent must be explicit and signed. Many journals require this as part of ethical approval before agreeing to publish research. (The same goes even for simple case-reports).

In order for this to happen, there must be *ethical approval*, and a detailed *protocol* which sets out what is intended to happen.

Ethical Approval

If using a SPOT to guide clinical care of a patient, ethics approval is not required on the assumption that the patient's best interests are served. A patient's implied consent is adequate.

To prepare an ethics approval application (which can be submitted to a university or professional college institutional ethics committee), there is a need to set out for both future patients and the committee what is intended. This will need to include all the features required by ethics committees (see Chapter 11).

Protocols

A formal protocol is not essential for use in therapeutic SPOTs (although it is a good idea to document what is intended in the patient's clinical record).

However for research, then a protocol is mandatory, if only for

1. The ethics application and
2. (In modified form as) an information sheet for the patient (as almost certainly required for ethics approval).

A written protocol specifies instructions for both doctor and patient, and allows ideas to develop greater clarity even when the SPOT will be done to assess the individual patient response. It may well additionally improve shared understanding.

The protocol should contain the following elements:

1. A brief literature search. This should detail any information about
2. What the best evidence shows about the efficacy and safety (associated adverse effects) of the intervention
3. What individual variation in response is documented
4. The method; what will be done? How? When? In particular think about each of the following:
5. Patients suitable for the SPOT
6. Treatment
7. Comparator
8. Outcome (what will be selected, how it will be measured, recorded and so on).

SPOT Example 1

"I want you to prescribe me testosterone."

The patient was hunched and ready for conflict from the consulting room chair.

"I have looked it up on the web," he continued, "and I am sure it will make me feel better".

"If you don't give it to me, I'll get something illegally," he added.

The GP re-checked his patient records. Of course he had no indication for supplemental testosterone. The GP began to gingerly explore the reasons for this 62 year old's odd request, suspecting some sexual problem, perhaps erectile dysfunction.

To his surprise the patient's concerns were not sexual. Rather they focussed on fatigue, or what he called 'energy levels'. He was not able to get out on the golf course as much as he used to, although he was able to keep on top of work (which was office work). Surfing on the Web he had found sites that suggested to him that testosterone would be effective.

The GP carefully went through his own misgivings. There was evidence that testosterone could invoke increased risk from a number of factors, principally thromboembolism. He would be prescribing it off-label for an unorthodox indication.

(continued)

SPOT Example 1 (continued)

On the other hand, the patient had clearly indicated that he was determined to obtain testosterone (or perhaps something even worse), and the GP thought he might do more good by supplying it and monitoring him.

So he proposed a SPOT. He agreed to prescribe the testosterone if the patient would agree to take it for alternate months (he said he thought it would 'work in a week or two'), and monitor some measure of 'energy level'. This took some negotiation. They both finally agreed on the number of:

1. times he entered the golf course per week;
2. days he felt energetic per week.

The patient entered the information into a diary, and both agreed he would return in 1 month for his first follow-up visit (after using the testosterone for a month).

At that consultation he said he had decided to give it up. It had made no difference to his energy levels, which was clear when he stood back to look at his diary (which was highly variable, anyway).

Whether or not this needed a SPOT to arrive at this decision is hard to know. The GP felt sure it avoided a confrontation between an ardent patient and reluctant prescriber, and the process of recording symptoms may have helped the patient realise that he was not going to dramatically improve his health with a single simple intervention.

SPOT Example 2

Jason was a 34 year old male accountant who came to see the GP for his chronic low back pain present since doing some heavy lifting over 3 years previously. Over this time the pain had spread up into his mid thoracic spine and out into both gluteal muscles. It was aggravated by activities such as lifting, ten-pin bowling and cycling but also by sitting and sleeping. After 8 h sleep he would wake with significant pain, stiffness and muscle spasm, causing great difficulties in straightening up. It took 3 h to ease in the morning with light activity. He got partial relief from osteopathy and massage and from over-the-counter anti-inflammatory medications taken intermittently. He had been investigated for an inflammatory cause with an ESR, CRP and HLA B27 genotyping, but all these tests were negative. A recent spinal MRI had shown early degenerative changes in his lumbar discs and facet joints, but no hallmarks of inflammation. There was a moderate restriction of lumbar flexion, but other spinal movements were normal

(continued)

SPOT Example 2 (continued)

and he had mild tenderness throughout his lumbar and lower thoracic spines, but not over his sacro-iliac joints. There were no peripheral signs of inflammatory joint disease.

Despite the negative investigations, the GP made a provisional diagnosis of sero-negative spondylo-arthropathy, and asked him to take 15 mg of meloxicam daily for 3 weeks. He returned enthusiastically reporting that he noticed considerable benefit after 1 week, with his morning pain reducing from 6/10 previously to 2–3/10. He had resumed roller-blading and was able to play with his children again without restriction. This seemed to support the GP's suspicion of inflammation, but given the implications of taking anti-inflammatory medication in the long term, the GP suggested a SPOT comparing meloxicam with no medication. Spurred on by his initial response and his analytical nature, Jason readily agreed to give this a try, even with the proviso that it would take 12 weeks to give a solid result. This was the timeframe for N-of −1 trials comparing celecoxib with paracetamol for osteoarthritis that the GP had previously managed in a research capacity. They had three 4-week cycles comprising blinded, randomised periods of 2 weeks on celecoxib and 2 weeks on paracetamol. Patients kept detailed diaries of five outcomes, some of which were daily, for the 12 weeks – quite an undertaking. The simplified adaptation of this involved alternating fortnights of 15 mg of meloxicam with fortnights of no medication, measuring average weekly morning and afternoon pain and stiffness at the end of each period. This protocol avoided the wash-in and wash-out periods for medication and made the recording processes less onerous than in an N-of-1 trial. The results are shown in Table 15.5.

Table 15.5 SPOT example 2 Low back pain and meloxicam

Time	2 weeks	2 weeks	2 weeks	2 weeks	2 weeks	2 weeks
	No medication	Meloxicam 15 mg/day	No medication	Meloxicam 15 mg daily	No medication	Meloxicam 15 mg/day
Symptom						
Pain on waking	7–8	4	5–6	3–4	7	4
Pain at 5 pm	3–4	1	2–3	0–1	3	1–2
Stiffness on waking	7–8	5	6	4	8	5
Stiffness at 5 pm	2–3	6	1–2	0–1	2–3	1

Doctor and patient discussed the pattern in his response of noticeably reduced pain and stiffness when on meloxicam with return of the same within several days of ceasing it.

Note that there was no statistical analysis – just a visual inspection of the results (Table 15.5) for patterns. Not captured by his recorded outcomes was an increased ability to exercise and play with his children during the periods on meloxicam. This was an important outcome to him and one, in retrospect, that should have been included. On the strength of his experience with this SPOT, Jason was keen to continue the meloxicam (under the cover of a proton pump inhibitor for gastric protection).

SPOT Example 3: Use of App

"I don't want to stop HRT because I'm worried that I won't sleep well without it."

Sandy was a 56 year old lady who had been on combined oestrogen/progesterone transdermal patches for 8 years. Treatment was originally started to relieve hot flushes and sweats. These had long since abated, but she thought that it helped with sleep too. Sandy was the single mother of two adolescent males and held a mentally demanding job with a high level of responsibility. She was concerned that her quality of sleep may deteriorate without the HRT.

The rest of the consultation did not uncover any other reasons why she may benefit from continued use of HRT.

Sandy was agreeable to a SPOT and to downloading the "Sleep Time" App which records duration of sleep, light and deep sleep, as well as sleep efficiency.

We agreed to 1 month of recording on treatment followed by 1 month of recording off treatment.

Contrary to expectations the App showed that the average duration and quality of sleep was not better on HRT (Table 15.6). As a consequence Sandy decided to stop using the HRT.

Table 15.6 SPOT sleep quality using HRT or not

Sleep – mean values	4 weeks on HRT	4 weeks off HRT
Duration	6.52 h	7.20 h
Ratio of deep to light sleep	80 %	82 %

Conclusion

In conclusion, what SPOT methodology lacks in rigor is compensated for by its ease of use in the workplace. SPOT research can never eliminate the placebo response. But SPOTs can arguably claim the top spot in patient centred research, both in treatment outcomes and patient consent. In this territory SPOT is a useful tool and placebo response is welcomed.

References

Barry MJ, Edgman-Levitan S (2012) Shared decision making – the pinnacle of patient-centered care. N Engl J Med 366:780–781

Bland JM, Altman DG (1994) Regression towards the mean. BMJ 308:1499

Brown M (2010) Charles Bridges-Webb AO 1934–2010. http://sydney.edu.au/medicine/alumni/news/tributes/100712.php. [Online]. University of Sydney. Accessed 28 Sept 2013

Copay AG, Subach BR, Glassman SD, Polly DW Jr, Schuler TC (2007) Understanding the minimum clinically important difference: a review of concepts and methods. Spine J 7:541–546

Howick J, Friedemann C, Tsakok M, Watson R, Tsakok T, Thomas J, Perera R, Fleming S, Heneghan C (2013) Are treatments more effective than placebos? A systematic review and meta-analysis. PLoS One 8:e62599

March L, Irwig L, Schwarz J, Simpson J, Chock C, Brooks P (1994) N-of-1 trials comparing nonsteroidal anti-inflammatory drug with paracetamol in osteoarthritis. BMJ 309:1041–1046

Stewart M (2005) Reflections on the doctor-patient relationship: from evidence and experience. Br J Gen Pract 55:793–801

Stewart MA (1995) Effective physician-patient communication and health outcomes: a review. Can Med Assoc J 152:1423–1433

Yelland MJ, Schluter PJ (2006) Defining worthwhile and desired responses to treatment of chronic low back pain. Pain Med 7:38–45

Chapter 16
Systematic Review and Meta-analysis Using N-of-1 Trials

Kerrie Mengersen, James M. McGree, and Christopher H. Schmid

Abstract This chapter discusses issues and approaches related to systematic review and meta-analysis of N-of-1 trials. Some basic guidelines and methods are described in this chapter. Some important steps in a systematic review of these types of trials are discussed in detail. This is followed by a detailed description of meta-analytic methods, spanning both frequentist and Bayesian techniques. A previously undertaken meta-analysis of a comparison of treatments for fibromyalgia syndrome is discussed with some sample size considerations. This is further elaborated on through a discussion on the statistical power of studies through a comparison of treatments for chronic pain. The chapter concludes with some final thoughts about the aggregation of evidence from individual N-of-1 trials.

Keywords Bayesian methods • Fixed effects models • Inclusion criteria • Meta-analysis • N-of-1 trials • Random effect models • Review question • Sample size • Statistical power • Systematic review

Introduction

An N-of-1 trial is a prospective observational study of a participant who is individually exposed to different treatments over time. It is also known as a single-patient trial, multiple crossover trial, or a form of single case design (Dallery and Raiff 2014) since the experimental evaluation occurs within the patient. The treatments applied to the patient are often randomized, replicated within the individual, and blinded. A principal attraction of N-of-1 trials is that they provide estimates of treatment effects on individuals, as opposed to average treatment effects obtained from

K. Mengersen (✉) • J.M. McGree
Mathematical Sciences, Queensland University of Technology, Brisbane, Australia
e-mail: k.mengersen@qut.edu.au; james.mcgree@qut.edu.au

C.H. Schmid
Department of Biostatistics and Center for Evidence Based Medicine, Brown University, Providence, RI, USA
e-mail: christopher_schmid@brown.edu

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_16

randomized controlled clinical trials, and have therefore been promoted as potentially useful for the individualization of medicine (Lillie et al. 2011; Duan et al. 2013). A comprehensive description of these types of trials is provided by Kravitz et al. (2014a) while Dallery et al. (2013) provide examples of other disciplines including psychology and occupational therapy where such designs have generated evidenced-based practices.

Although a trial can comprise a single individual, they commonly comprise multiple individuals. The multiple subjects designs allow for evaluation and comparison of treatments both within and across individuals. That is, the data from the individuals can be statistically combined to provide individual treatment-effect estimates which 'borrow strength' across other similar patients, and also provide average treatment effects. In a similar manner, evidence can be combined among different N-of-1 trials conducted with different groups of patients. These combinations can be performed through systematic reviews and meta-analysis.

Systematic reviews provide a framework for consistent evaluation of studies undertaken for the purpose of addressing a common scientific question, where 'scientific' is used here in a broad sense and covers medicine, science, social science, environment and ecology, finance and economics, and so on. Systematic reviews are endemic in medical research and are often a compulsory component of evidence-based medicine. They are also becoming standard practice in other fields; see, for example, the guidelines for systematic reviews and meta-analysis in ecology and evolution detailed in Koricheva et al. (2013).

Meta-analysis is the quantitative combination of statistical estimates from a collection of studies, where these are often compiled as part of a systematic review. The methodology employed for meta-analysis depends on a range of statistical considerations as well as the aim of the meta-analysis itself. There is a large literature on meta-analysis, particularly in the field of clinical medicine and health. The purpose of this chapter is to discuss the way in which evidence from N-of-1 trials can be used for systematic reviews and meta-analysis. The process of conducting a systematic review is discussed in section "Systematic Reviews of N-of-1 Trials" and of undertaking a meta-analysis in section "Meta-analysis of N-of-1 Trials". Sample size considerations are discussed in section "Meta-analysis Modelling Decisions". The chapter concludes with a general discussion and directions for future research.

Systematic Reviews of N-of-1 Trials

No matter how well a systematic review or meta-analysis is conducted, poor quality evidence in the individual trials included in the meta-analysis leads to interpretations and conclusions that are at best unable to say anything of value, and at worst highly misleading. The process of undertaking a systematic review has been well documented for clinical studies through the Cochrane Collaboration; see, for example, the Cochrane Handbook Systematic Reviews of Interventions (Higgins and Green 2008) which provide detailed recommendations on procedures for defining

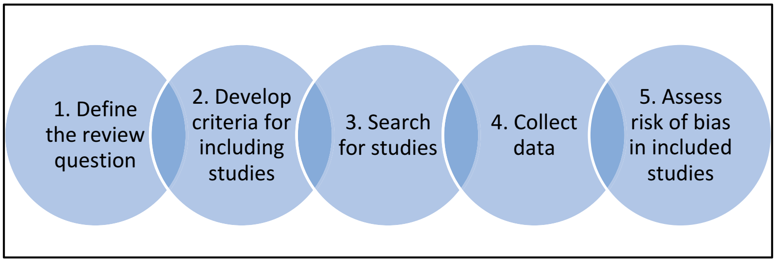

1. Define the review question
2. Develop criteria for including studies
3. Search for studies
4. Collect data
5. Assess risk of bias in included studies

Fig. 16.1 Selected steps in a systematic review

the review question, developing criteria for including studies, searching for studies, collecting data, and assessing risk of bias in included studies, among other issues. These steps are depicted in Fig. 16.1. Other fields have also developed guidelines; see, for example, Chaps. 4 and 5 in Koricheva et al. (2013) which deal with searching literature, criteria for selection of studies, and extraction and critical appraisal of data.

Below we discuss the application of the first two steps depicted in Fig. 16.1, which are most relevant for systematic reviews of N-of-1 trials.

Define the Review Question

As with all meta-analyses, the review question must be clearly and specifically defined. It must be sufficiently specific to allow the combined results to be interpretable, yet sufficiently broad to allow for a sufficient number of studies to be included in the analysis.

In formulating the review question, one can reflect upon the suggestions of Zucker et al. (2010) in regards to the type of studies N-of-1 trials are best applied to. These are:

- The condition must be chronic and stable.
- The interventions must be symptomatic (not permanently changing the condition status).
- The interventions need to have appropriate on/off kinetics to limit possible carryover and period effects.

Develop Criteria for Including Studies

Because N-of-1 trials are individually tailored to a single patient, the criteria for inclusion of studies will to some extent be specific to the problem. However, there are common considerations that will apply to most, if not all, reviews:

Table 16.1 Five key characteristics in assessing N-of-1 studies

Key characteristics	
1. *Why was an N-of-1 design chosen?*	
What was the motivation for this study?	√
Why was this design selected?	√
2. *Was the trial well designed?*	
Was the design appropriate for an N-of-1 trial?	√
Were other potential biases and confounders addressed in the design?	√
3. *Was the study appropriately conducted?*	
Was the study conducted according to the design?	√
What potential biases and confounders were induced in the conduct of the study, and how were these addressed?	√
4. *Was the trial correctly analyzed?*	
Were appropriate statistical methods used for the analysis of the data?	√
Were potential biases and confounders addressed in the statistical analysis?	√
Was the statistical analysis sufficiently comprehensive and interpretable?	√
5. *Were the analyses correctly reported?*	
Was the statistical analysis adequately reported?	√
Were the conclusions of the study appropriate given the study design, conduct and analysis?	√

1. The studies must be relevant to the review question.
2. The studies must be of sufficiently high quality to merit inclusion in the review.
3. The studies must report sufficient information to enable key characteristics to be extracted for the review.

Steps 1 and 3 are relatively straightforward. Step 2 requires consideration of five key characteristics of N-of-1 trial quality, as indicated in Table 16.1.

A brief discussion of these considerations is given below. Further details are found in other chapters in this book.

Why Was an N-of-1 Design Chosen?

What Was the Motivation for This Study?

A scientific study is often proposed when there is substantial uncertainty about the answer to a question of interest. Clinical trials are often proposed when the question of interest is a comparison of the effectiveness of specified treatments. Most trials are constructed to learn about the average treatment effect across a group of individuals receiving a treatment. An N-of-1 trial is often proposed when interest is focused on the efficacy of treatment for a single individual, in order to make a clinical decision.

It is useful to understand the background and motivation of the trial. For example, is the scientific study motivated by a general lack of knowledge, or because there is conflict regarding existing evidence, or because the available evidence is possibly not relevant to the particular question? Is the clinical trial based on a set of well defined treatments, and are the measures used in the study clinically accepted?

It is useful to learn who was involved in developing the study. Kravitz et al. (2014b) argues that a successful N-of-1 trial requires a close collaboration between the patient and the clinician.

Why Was This Design Selected?

A study should provide justification of the use of this design as opposed to alternatives, such as the gold-standard randomized parallel group trial.

Kravitz et al. (2014b) provide a detailed exposition of the reasons why such a design might be selected. In summary, an N-of-1 trial is most useful:

- When there is substantial variation in treatment outcomes within the individual, and when interest focuses on these treatment outcomes for the individual;
- If the variation in treatment effects across patients cannot be easily predicted from available prognostic factors but is anticipated to be substantial;
- If the outcome of interest is chronic, stable or slowly progressive and is either symptomatic or is associated with a valid biomarker;
- If the outcome is rare, so that there is little other evidence of treatment effect;
- If the treatments have relatively rapid onset and washout; and
- If the treatment regime is relatively straightforward.

Was the Trial Well Designed?

Was the Design Appropriate for an N-of-1 Trial?

The design of an N-of-1 trial requires careful consideration (Kravitz et al. 2014b; Schmid and Duan 2014). These trials are subject to usual study design issues such as randomization, replication, blocking, the choice of outcomes, the scale of the outcomes (continuous, categorical or count data). They are also subject to other specific issues, including the design of crossovers, time-dependent confounders and changes over time independent of treatment, carryover of treatment effects from one period into the next, auto-correlation of measurements and premature end-of-treatment periods.

Kravitz et al. (2014b) describe five important characteristics of a well-designed N-of-1 trial: balanced sequence assignment, repetition, washout and run-in, blinding, and systematic outcomes measurement. Schmid and Duan (2014) reiterate these design principles, identifying in particular four considerations: randomization and counterbalancing, replication and blocking, the number of crossovers needed to optimize statistical power, adaptation, and the choice of outcomes of interest to the patient and clinician.

Balanced sequence assignment aims to ensure that the estimates of treatment effects within an individual will not affected by time-dependent confounders. This can be achieved by randomization of treatment periods or by careful experimental design. While randomization aims to achieve balance when averaged over a large number of blocks, individuals or trials, *counterbalancing* aims to achieve exact or nearly exact balance in each individual. A counterbalanced design takes the form of an ABBA or BAAB sequence of treatments so that the treatments do not always appear in the same order (e.g., ABAB) and are not related to a time sequence (e.g., AABB).

Replication of treatments ensures that treatment effects are not confounded with other factors. These factors can be due to the individual, such as random changes in diet, the environment, such as changes in weather, or treatment-related, such as random variation in the outcomes being measured in the trial. The number of measurements taken on an individual depends on the number of treatment periods, the length of each period and the frequency of measurements per period. These choices, and the corresponding allocation of measurements, will depend on practical considerations, but it should also ensure that adequate estimates of within- and between-treatment variation (and between-individual variation in multiple N-of-1 trials) can be obtained for the statistical analysis. *Blocking* is a form of repetition or replication within the individual, with a systematic allocation of treatments chosen to protect against, systematic, time-dependent and random variation. Treatments are randomized or counterbalanced in small groups such as of size two or four. The additional balance induced by blocking also reduces adverse consequences of early termination from the trial.

Overall, Schmid and Duan (2014) recommend a blocked design for N-of-1 trials, potentially combined with counterbalancing if there is good information on the most important potential confounding factors (such as the linear time trend) and if the total number of blocks is small (e.g. less than four). Otherwise blocking with randomization is recommended.

A *wash-out* period separates the active treatment periods in an individual, and can be used to mitigate the effect of sequential treatments. A run-in period occurs between enrolment of the individual and randomization of the treatments and can be used to monitor patients' baseline outcomes. The pros and cons of washout periods are discussed by Kravitz et al. (2014a).

A *blinded* study prevents the clinician, subject (patient, participant), and/or evaluator (person taking the measurements) from knowing the assignment sequence. This is an established scientific study protocol aimed at preventing potential conscious or unconscious bias. Schmid and Duan (2014) describe various reasons why bias might arise in these types of studies. Although blinding is not always feasible, studies that do not adopt it should carefully justify their choice and describe how potential biases were mitigated and/or accommodated in the statistical analysis.

An *adaptive* trial allows the design to be modified during the course of the trial to improve efficiency or resolve problems in its design or conduct. Adaptive trials are now well accepted and have been demonstrated to be very powerful in detecting treatment effects and comparisons more quickly and efficiently. However, the adaptation rules need to be developed before the trial commences and potential biases need to be mitigated by strategies such as blinding.

Finally, *systematic outcomes measurement* refers to the identification of what data to collect and how to collect them. Kravitz et al. (2014b) describe the process of identifying outcome domains, indicators or measures of those domains, and data collection. Whereas other clinical trial designs typically select a primary measure of interest and design the trial around this, N-of-1 trials focus on the outcomes of importance to the individual patient and his or her clinician. These will often be analyzed separately, but can be combined in some form of index or composite measure. While this offers substantial flexibility and the possibility of more effectively answering the aims of the trial stakeholders, it is important to verify the acceptability, reliability, validity and relevance of selected measures. Importantly, do they provide comprehensive coverage of the question of interest and do they measure what they are intended to measure? A range of data types and sources, ranging from surveys to mobile devices, is potentially available for use in N-in-1 trials. The chosen observations, which will form the trial statistics, must be adequate estimators of the measure of interest, both clinically and statistically. A preferred estimator, and corresponding statistic, is one that is accepted in the literature and is unbiased and consistent. That is, it can used for comparative purposes in a systematic review or meta-analysis, if it accurately estimates the target measure and the precision of the estimate improves as the sample size increases.

An extract of the checklist developed by Schmid and Duan (2014) for assessing the design of an N-of-1 trial is reproduced in Table 16.2.

Were Other Potential Biases and Confounders Addressed in the Design?

The design principles described above are intended to avoid a range of well-recognized potential biases and confounders in an N-of-1 trial. Other biases and confounders can arise in these studies. Often these are situation-specific. Consideration of potential issues in the trial design that may impact on the statistical estimates is important if these trials and corresponding estimates are to be included in systematic reviews and meta-analyses.

Was the Study Appropriately Conducted?

Was the Study Conducted According to the Design?

The typical N-of-1 trial is based on an individual design developed between trial stakeholders, typically a clinician and a patient. The stakeholders therefore have an invested interest in conducting the study according to the agreed design: it is likely to be practically achievable and to answer questions of direct interest. However, there are always variations between intent and execution, i.e. between trial design and conduct. It is important that these differences are documented.

Table 16.2 Checklist for design of N-of-1 trials (Adapted from Schmid and Duan (2014))

Guideline	Consideration
Balance treatment assignment across conditions, using either randomization or counterbalancing, along with blocking	Design needs to eliminate or mitigate potential confounding effects such as time trend
	Pros and cons of randomization versus counterbalancing need to be considered carefully and selected appropriately. Counterbalancing is more effective if there is good information on a critical confounding effect, for example, linear time trend. Randomization is more robust against unknown sources of confounding
	Blocking helps mitigate potential confounding with time trend, especially when early termination occurs
Blind treatment assignment when feasible	Blinding of patients and clinicians, to the extent feasible, is particularly important for N-of-1 trials, especially with self-reported outcomes, when it is deemed necessary to eliminate or mitigate nonspecific effects ancillary to treatment
	Some nonspecific effects might continue beyond the end of trial within the individual patient, and therefore should be considered part of the treatment effect instead of a source of confounding
Use appropriate measures to deal with potential bias due to carryover and slow onset effects	A washout period is commonly used to mitigate carryover effect. Adverse interaction among treatments being compared indicates the need for a washout period
	Absence of active treatment during a washout period might pose an ethical dilemma and diminish user acceptance for active control trials
	Washout does not deal with slow onset of new treatment and might actually extend duration of transition between treatment conditions
	Analytic methods can be useful for dealing with carryover and slow onset effects when repeated assessments are available
Perform multiple assessments within treatment periods	This increases the precision of estimated treatment effect and facilitates analytic approaches to address carryover or slow onset effects
	The cost and respondent burden need to be taken into consideration in decisions regarding frequency of assessments
Consider adaptive trial designs and sequential stopping rules	These can help improve trial efficiency and reduce patients' exposure to the inferior treatment condition

One important issue raised by Schmid and Duan (2014) is data collection. While the motivation of patients to participate in an N-of-1 trial may reduce the problem of missing data, the complexity of the design, multiplicity of outcome measures and lack of easy access to standard data acquisition tools such as forms and software can increase the problem. This needs more careful attention in these types of trials than in more standard randomized controlled trials for which the infrastructure, data collection rules and mechanisms, and trial support are more widely established.

What Potential Biases and Confounders Were Induced in the Conduct of the Study, and How Were These Addressed?

Unplanned and unexpected events during the course of a trial can induce potential biases and confounders. Examples of these include unexpected termination of the study, missing data not at random, unplanned changes in patient characteristics relevant to the study and unplanned external influences on the trial. These need to be carefully evaluated with respect to their potential influence on the trial analysis and results. If the impact is substantial, the trial should be adapted or terminated.

Was the Trial Correctly Analyzed?

Were Appropriate Statistical Methods Used for the Analysis of the Data?

Most N-of-1 trials published to date have used graphical comparisons, a statistical cutoff or a clinical significance cutoff to compare treatments (Gabler et al. 2011). Graphical and statistical summaries are always helpful in the preliminary analysis of data, and they can often facilitate inferences if the study design is simple and the results are clear. However, models are often important.

Schmid and Duan (2014) provides details of the statistical methods used for analysis of individual participant data from N-of-1 trials. These methods are represented as a decision tree in Fig. 12.1.

Were Potential Biases and Confounders Addressed in the Statistical Analysis?

As indicated in Fig. 12.1, many analyses ignore both time-related effects and the fact that the measurements in an N-of-1 trial are correlated. Correlation between measurements within periods, and carryover effects between treatment periods, can be accounted for using established time series methods, such as autoregressive models and dynamic models. The models can also account for time. Depending on the nature of the time dependence, this can be achieved by indexing time within treatment period and/or by indexing treatment periods within blocks.

Was the Statistical Analysis Sufficiently Comprehensive and Interpretable?

As argued by Schmid and Duan (2014), a Bayesian model is better able to describe the complexities of an N-of-1 trial described above, compared with a standard frequentist model, because it is more modular in structure, can include prior information about treatment differences or measurement errors and biases, and can incorporate different sources and types of information more easily. Moreover, a Bayesian analysis provides more valuable inferences than a standard frequentist analysis, because the posterior probability resulting from the Bayesian analysis is more interpretable and able to deliver a richer set of results than the p-value that is typically reported from a frequentist analysis. The posterior probability allows stakeholders to obtain a wide range of relevant estimates, comparisons and probabilities, with corresponding statements about the uncertainty of these values. These can be for a composite outcome value or for multiple outcomes, where the latter are described by joint posterior probability distributions. Schmid and Duan (Schmid and Duan 2014) provide examples of the types of outputs and inferences that can be obtained from a Bayesian analysis, including the probability that one treatment is superior to another treatment, probabilistic ranking of multiple treatments, the probability that the treatment effect is at least as large as a certain clinically important size, and so on.

A Bayesian analysis of the individual N-of-1 trial also allows a more streamlined analysis of multiple N-of-1 trials, as described in section "Meta-analysis Modelling Decisions".

Was the Trial Correctly Reported?

Was the Statistical Analysis Adequately Reported?

Adequate reporting of statistical results is an acknowledged problem in most studies and can arise because of ignorance on the part of the authors, pressure for space on the part of the journal or reporting agency (although this is less defensible with the increasing adoption and availability of online supplementary material), lack of systematic reporting requirements and different intended uses of the reported information. In the systematic review reported by Gabler et al. (2011), most of the 108 trials reporting on 2154 subjects provided at least some relevant quantitative information (e.g. percentages) related to the treatments, but less than half of the trials (45 %) reported adequate information to facilitate the statistical estimation of treatment effects.

Were the Conclusions of the Study Appropriate Given the Study Design, Conduct and Analysis?

The adequacy of the reporting of statistical results can pose a problem when evaluating studies. First, it can sometimes be difficult to decide whether results are presented correctly because the required information is not presented, as described above. Second, the reported information may be limited (e.g., percentages), requiring substantial interpretation by the author of the report which may be difficult to verify.

Reporting of conclusions is a third, related issue: the relevant information may be presented, but if it is not compelling then an enthusiastic author might make progressively more assertive statements from the 'Results' section to the 'Conclusions' section and from there to the 'Abstract'. If an unsuspecting systematic reviewer selects studies, develops summaries or makes inferences based on published abstracts, there is a strong potential for erroneous conclusions.

Meta-analysis of N-of-1 Trials

Meta-analysis of N-of-1 trials occurs at three levels:

1. Combination of treatment outcomes within an individual; this is described in section “Systematic Reviews of N-of-1 Trials” above
2. Combination of results for a number of individuals in a multi-treatment study
3. Combination of results from a number of multi-treatment trials.

The first level is described in section “Systematic Reviews of N-of-1 Trials” above. The second level, and to some extent the third level, are described by Zucker et al. (1997), Zucker et al. (2010), Duan et al. (2013) and Schmid and Duan (2014). Other researchers undertake the analysis of individual trials (first level), as well as the synthesis of results across many individuals (second level), see Senior et al. (2013). A summary of the methods proposed for the second and third levels is given here.

Should the Trials be Combined?

Prior to meta-analysis, it is important to ask whether it is scientifically and clinically valid to combine the trials. Results of individual N-of-1 trials may be combined if the trials are considered to be sufficiently similar with respect to the trial administration (e.g. the same clinician), the treatments administered in the trials, the characteristics of the individual patients, the trial design, and so on. The combined results must be interpretable in some way: they must give useful information about treatment comparisons and about the cohort of subjects to whom the comparisons are applied. If there is too much variability or uncertainty in any of these factors, the trials should not be combined.

If it is determined that the trials are sufficiently similar to warrant combination, then a variety of statistical techniques can be applied, depending on the data. Table 16.3 provides a summary of the methods identified by Zucker et al. (2010).

Statistical Models for Meta-analysis

Zucker et al. (2010) provided details of the models, assusmptions and inferences for the approaches for combining all data available from the N-of-1 trials and compared these with approaches using summaries or portions of the data. The models are briefly presented here; see Zucker et al.(2010) for explanation and discussion.

Table 16.3 Meta-analysis models for N-of-1 trials, extracted from Zucker et al. (2010)

Type of data	Types of models
Data aggregated to the trial (patient) level	Summary fixed and random effects models
Data at trial-period level: multiple estimates of an effect per trial	Summary random-effects model or mixed model
Subset of prospective data with treatment order randomized across trials, e.g.: (i) first period treatments, analogous to a randomized parallel group trial; (ii) pair-randomized treatments in first two periods (AB/BA crossover design)	Standard model for analysis of population designs, e.g. (i) t-test; (ii) paired t-test
Data using all periods	Fixed or random effects model; Multiple crossover model; Repeated measures model; Linear mixed model; Bayesian hierarchical model

Summary Fixed and Random Effects Models

Assume that there is a summary effect y_i from each trial. This could be a single outcome measure or a composite measure. In a fixed effects model, these are assumed to vary randomly around an overall true mean effect α:

$$y_i = \alpha + \varepsilon_i\,;\ .\varepsilon_i \sim \mathrm{N}\left(0, \sigma_i^{\ 2}\right)$$

where σ_i^2 is the variance associated with y_i. If there is no repetition within a trial, these variances will not be available; if the trials have similar designs then the variances can be assumed to be equal and the analysis can proceed. Otherwise alternative assumptions have to be made. If there is only a small number of repetitions or treatment periods per trial, so that each trial variance is poorly estimated, it may be preferable to replace σ_i^2 by a common pooled variance σ^2. For N-of-1 studies, such variances are often assumed known, but this may be problematic if the number of observations is small (usually would not have no replication in an N-of-1 study). See Zucker et al. (2010) for details.

If there are multiple outcome measures, the model becomes multivariate, with y_i becoming a vector of estimated effects from the ith study, α becoming a vector of mean effects, and ε_i having a multivariate normal distribution with $\sigma_i{}^{2}$ replaced by a variance-covariance matrix Σ_i.

In a random effects model the estimated effect y_i is assumed to vary around a trial-specific effect $\alpha_{i,}$ which is in turn assumed to vary around an overall effect α_0:

$$y_i = \alpha_i + \varepsilon_i;\ .\ \varepsilon_i \sim \mathrm{N}\left(0, \sigma_i^{\ 2}\right);\ \alpha_i \sim \mathrm{N}\left(\alpha_0, \tau^2\right)$$

or alternatively and equally

$$y_i = \alpha_0 + \alpha_i + \varepsilon_i;\ .\varepsilon_i \sim \mathrm{N}\left(0, \sigma_i^{\ 2}\right);\ \ \alpha_i \sim \mathrm{N}\left(0, \tau^2\right).$$

Here, σ_i^2 describes the variation of the effects within a trial (the 'within-trial variance' and τ^2 describes the variation of effects among or between trial (the 'between-trial variance'). Note that there need to be enough trials to adequately estimate the between-trial variance τ^2. If this is not the case, a fixed effects model might be preferred. See Zucker et al. (2010) for further discussion of this issue.

Mixed Models

Mixed models aim to combine within- and between-patient data simultaneously, as in an individual patient data [IPD] meta-analysis. Let y_{ij} be the observed effect for the *i*th trial (patient) in the *j*th period, $j = 1,..,J$, and let $\underline{y}_i = (y_{i1},..,y_{iJ})$. Then the mixed model is simply a variation of the random effects model described above: $\underline{y}_i$ is assumed to have a multivariate normal distribution with mean $\underline{\mu}_i$ and $J \times J$ variance-covariance matrix Σ_i. The trial design and sources of variation within and among the trials may dictate the way in which $\underline{\mu}_i$ and Σ_i are defined. Different definitions give rise to the multiple crossover and repeated measures models listed in Table 16.2. Detailed examples are given in Zucker et al. (2010).

Modelling the Complete Dataset

The above models can be extended to analyze the full set of data available from multiple N-of-1 trials. As described by Schmid and Duan (2014), let *m, i, j, k* and *l* denote the individual (trial), observation, treatment period, block and treatment, respectively. Then a simple random-effects model for observation y_{mijkl} is given by

$$y_{mijkl} = \alpha_m + \beta_l + \gamma_k + \delta_{j(k)} + \varepsilon_{i(j(k(m)))}$$

where the four terms indicate the variability among individuals, treatments, blocks, treatment periods within a block, and observations within a treatment period within a block within a patient. The treatment effect is considered fixed; the individual or trial is considered to be random with distribution $\alpha_m \sim N(\alpha_0,\sigma_a^2)$ where α_0 is the overall effect, and the other three effects are also considered to be normally distributed with means 0 and variances σ_γ^2, σ_δ^2 and σ_ε^2, respectively.

Allowing for Time-Related Effects

The above models can also be extended to allow for time trend and carryover. As described by Schmid and Duan (2014), a meta-analysis model for outcome *y* for the *i*th patient that incorporates a time trend at time *t* is given by

$$y_{it} = \alpha_i + \beta_t T_r + \gamma X_t + \varepsilon_{it}$$

where T_t is the time at time t and X_t is an indicator for the treatment received. This model returns an estimate of the trial effect (i.e. individual effect, given by α_i), the linear trend over time (given by β), the treatment effect (given by γ) and the residual variation (given by ε_{it}).

These models can be extended in a straightforward manner to include correlations in the residuals over time, nonlinear terms to capture possible nonlinear trends, seasonal effects, interactions between patients and other factors explaining variation across patients.

Bayesian Models

Bayesian models build on the above formulations by adding priors to each of the unknown parameters and expressing the parameter estimates in the form of posterior distributions (instead of maximum likelihood estimates as in frequentist analyses). Inferences of interest, such as comparisons and rankings of treatment effects, probabilities that treatment effects exceed thresholds of interest, etc. are then derived from these posterior distributions. See Zucker et al. (1997, 2010), Duan et al. (2013), and Schmid and Duan (2014) for more detailed explanations and examples of Bayesian approaches.

Meta-analysis Modelling Decisions

Zucker et al. (2010) describe a case study in which they combined 58 N-of-1 trials comparing amitriptyline (AMT) and the combination of AMT and fluoxetine (FL) for treating fibromyalgia syndrome (FMS). Details of the study are provided by Zucker et al. (2006). The trials had the following characteristics:

- Each trial had six treatment periods: three sets of paired treatments, comprising one period on AMT + FL and one on AMT.
- All treatment pairs were block randomized.
- The outcome measure (the quality-of-life Fibromyalgia Impact Questionnaire (FIQ) score) – a continuous value between 0 (best) and 100 (worst) was measured prior to any FMS medications and again at the end of each of the six 6-week treatment periods.

Zucker et al. (2010) analyzed data from the 46 patients who completed at least one period on each treatment. Of these patients, 34 completed all six treatment periods. The authors illustrated the application of a range of methods for meta-analysis. In particular, a variety of mixed models were fitted to the aggregated data, and the reader is encouraged to review their work. The meta-analysis models differed in how the intercept and treatment effect were treated (fixed and/or random), how patients' variances were treated (equal or unequal variances) and how the within-patient variance was structured (for example, single and uncorrelated). They made a number of comments regarding the implications of sample size in these analyses.

Choosing Between Fixed Effects and Random Effects Meta-analyses

In choosing between a fixed effects meta-analysis and a random effects meta-analysis, choose a random effects approach if there is substantial variation between trials compared to within trials, and/or to reduce the sensitivity to large differences in within-trial variances.

This is not a unique feature of N-of-1 trials but is a common phenomenon in all meta-analyses. Whereas a fixed effects meta-analysis includes only within-trial variances, a random effects meta-analysis includes both within- and between-trial variances. The overall effect is calculated as a weighted average of each of the trial-specific effects; under the fixed effects model these weights only involve the (inverse of the) within-trial variances, and under the random effects model the weights involve the combination of the within- and between-trial variances. Thus for a random-effects model, as the variation between trials increases, the relative influence of the within-trial variances decreases and the trial weights become more similar.

Robust Estimation Can Be Further Increased by Using a Common Estimate of the Within-Trial Variance

The variance of estimated effects for a typical N-of-1 trial is usually poorly estimated, since the sample size is usually small. For the simple fixed and random effects models, a common within-trial variance can be calculated by pooling across trials. Similarly, for a mixed model (e.g., nesting sets of treatments within treatment periods in the case study described above), a common within-trial covariance matrix can be calculated. In addition to providing a more robust estimator of the within-trial variance, this approach also reduces the number of parameters that need to be estimated. For example, in the above case study, the number of parameters in a mixed model meta-analysis can be reduced from six variance and 21 covariance terms in a full model (with different within-trial variances and covariances) to one variance term (assuming common within-trial variances and uncorrelated trials). Note that different model assumptions can be considered and evaluated with respect to stability of estimation and interpretability of results.

In a Bayesian Analysis, the Use of Appropriate Prior Information Can Improve Estimates

In the above case study, Zucker et al. (2010) derived prior distributions from a published crossover trial that used the same medications and dosages. They showed that this produced more robust estimates, in the sense that they were not only based on the small number of available observations, but also facilitated the estimation of otherwise unavailable parameters such as trial-specific variances and covariances. One difference between a meta-analysis of N-of-1 trials and that of randomized trials is that the number of trials in an N-of-1 study is usually substantially more than the number available from clinical trials, so that the data provide more information about the between-trial variance. The Bayesian model is therefore less sensitive to the prior on this parameter.

Sample Size Considerations in Meta-analyses of N-of-1 Trials: Trial Precision

The models described in section "Meta-analysis of N-of-1 Trials" and the considerations listed above highlight the importance of being able to accurately estimate the within-trial and between-trial variances. This necessarily depends on the number of treatments per trial and the number of trials.

How Should These Numbers Be Chosen?

Duan et al. (2013) studied this question by adopting a simple random effects model and calculating the variance of the mean effect with M trials and N paired treatment periods per trial, compared with a classic two-period (AB/BA) crossover design under several combinations of values of the between-trial variance (τ^2) and within-trial variances (σ^2/N).

Assuming independent trials, the calculated precision is $\omega = M / \left(\tau^2 + 2\sigma^2 / N\right)$. The following observations were made by the authors.

- For fixed τ^2 and σ^2, the value of ω increases as the number of trials (M) increases and as the number of repeated measures within a trial (N) increases.
- The relative importance of M and N depends on the relative size of the within- and between-trial variances.
- Additional measurements on individual patients are valuable if the between-trial variability is small compared with the within-trial variability.
- Conversely, more trials are more valuable than more measurements on individuals if the within-trial variance is small compared with the between-trial variance.

The effect on ω of M, N, τ^2 and σ^2 is illustrated in Fig. 16.2. Here, the horizontal and vertical axes show values of the within- and between-trial variances, respectively, and the four plots show different combinations of the trial size and number of trials. The contour lines show the value of the variance $1/\omega$. Comparison of the orientation of the contours and their relative magnitude indeed reveals and supports the above observations.

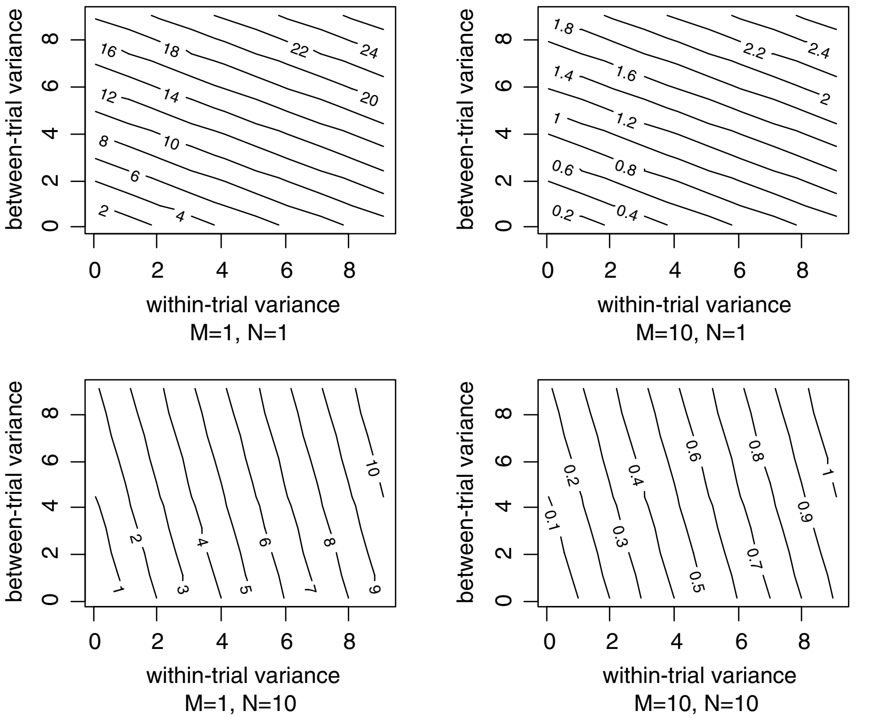


Fig. 16.2 Contour plots of the value of $1/\omega = 1/\left\{M/\left(\tau^2 + 2\sigma^2/N\right)\right\}$ for different values of M, N, τ^2, σ^2

These principles can be applied even when more complex meta-analysis models are employed. Alternatively, more refined calculations can be made for these models by examining the role of the within- and between-trial variances in the relevant equations for the variance components. As a general rule, more measurements should be taken of parameters which are poorly estimated, but this will depend on the accuracy required for the overall estimates and the role that the variances play in these estimates.

Further Sample Size Considerations in Meta-analyses of N-of-1 Trials – Statistical Power

The above discussion of sample size can be extended to meta-analyses of N-of-1 trials for the comparison of treatments through the consideration of statistical power. This can be defined as the probability of detecting a difference between treatment effects, given that a certain difference actually exists. In many instances, some prior knowledge, whether it be from previous studies or expert opinion, about the true difference or the clinically relevant difference between treatment effects can be obtained. This can then be used in formulating a study with high power.

We consider statistical power of a meta-analysis through a previously conducted N-of-1 randomized trial for the assessment of the efficacy of Gabapentin over placebo for chronic neuropathic pain (Yelland et al. 2009). Details of the study can be found in the reference, but let's suppose we are interested in conducting a similar meta-analysis, with functional limitation as the primary outcome. Based on the Hierarchical Bayesian meta-analysis conducted by Yelland et al. (2009), it was estimated that the difference between treatment effects was 0.6 (0.2). If we fix the number of paired cycles per individual at six, then the question is how many individual trials are required to maintain a high probability of determining a difference between treatments assuming a difference in effects of 0.6 exists?

To answer this, we must first be clear about how the treatment difference will be estimated (that is, how the data will be analyzed). Here, let's assume that we follow the methodology above in '*Summary fixed and random effects models*' and fit the specified random effects model with no block, period or order effects. Further we assume that the individual effects and residual variability follow a normal distribution each with variance of one. It is also assumed that patients have equal, uncorrelated response variances and equal variances by treatment. Of course, uncertainty in the parameters (example, the standard error of 0.2 for the difference in treatment effects) and model/s (for example, equal variances by treatment or the inclusion of a block effect) can be included but this was not thought to be relevant for this chapter. It is important to note that the estimates of power will depend on such assumptions.

With the analysis plan clearly specified and relevant parameters defined, statistical power can be estimated. In this work, we simply estimated power via simulation. That is, initially we simulated patient data from the assumed model, re-fit the model to the simulated data, conducted an hypothesis test to determine if there was a significant difference between treatments (significance level used here was 0.05), recorded the result of the hypothesis test, then repeated the whole process a large number of times (here we chose to repeat the process 500 times). The proportion of times the null hypothesis was rejected is the estimate of statistical power. This estimate is shown in Fig. 16.3 for a variety of different numbers of patients.

From Fig. 16.3, the power of the study increases as the number of subjects increases. In general, it is thought that 80 % power is reasonable, and it appears that this would be achieved with about 22 individual trials. This can be improved to about 90 % with an additional 11 trials. In estimating statistical power, simulation techniques were used to mimic the potential data which might be observed in the meta-analysis. An important part of this data simulation was to allow for the occurrence of missing data as this has the potential to significantly reduce the power of the study. For example, from Yelland et al. (2009), only 75 % of individual trials yielded at least one cycle, and only 65 % of trials yielded all three cycles. From these percentages, it is clear that such trials can be subject to many missing data points, and this should be accounted for in the simulation when estimating statistical power.

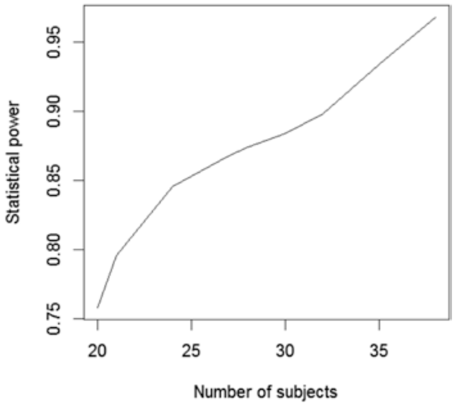


Fig. 16.3 Estimated statistical power for different numbers of patients for a hypothetical N-of-1 trial of Gabapentin versus placebo

Discussion

This chapter has described the conditions under which evidence from N-of-1 trials can be included in a systematic review or meta-analysis. The overall answer is *yes, with reservation.*

The first reservation is that the N-of-1 trials themselves need to be carefully and well designed. As discussed in section "Systematic Reviews of N-of-1 Trials", the quality of these trials is a paramount issue for systematic reviews and meta-analyses. Schmid and Duan (2014) argued that although N-of-1 trials allow great flexibility in meeting the aims of the patient and clinician and conforming to individual constraints, they also need to adhere to good design principles if they are to deliver accurate, replicable and comparable evidence. They suggest that a centralized service responsible for designing these trials might assist clinicians who are unfamiliar with these principles and hence ensure that proper standards are maintained, while still allowing designs to remain flexible and easy to implement.

The second reservation is that the systematic review must be designed, conducted and reported in such a way that it facilitates a systematic comparison while allowing for the individual characteristics of the trials. The systematic review reported by Gabler et al. (2011), based on 2154 participants in 108 studies published between 1985 and December 2010, found that N-of-1 trials were useful for increasing precision of estimates for a range of medical conditions, but recommended that the trial results include a clear description of individual data in order to facilitate future meta-analysis. Extending this observation, the 'clear description' should comprise common components that enable the systematic comparison to be undertaken. This is achievable. While the guidelines provided by the Cochrane Collaboration may be the gold standard for randomized controlled clinical trials, there are also parallels for less well designed studies, such as those developed by the Centre for Evidence Based Conservation and National Centre for Ecological Analysis and Synthesis in ecology and the Campbell Collaboration in the social sciences.

The systematic review must also be representative in some sense, in that the comparisons and generalizations arising from the review are applicable to a recognized population. If the review only contains trials that are published because the treatment comparisons are significant (so-called publication bias or the file-drawer problem), then the generalizations will not be applicable to the whole population. Notwithstanding this, a systematic review may still be useful for noting strengths and weaknesses of the trials, issues in reporting, deficiencies in publication or access to trial information, other issues related to systematic comparisons and other information gaps.

The third reservation is that a meta-analysis based on studies that vary substantially with respect to design and reporting needs to be very carefully formulated. The statistical model needs to accommodate these variations in order to deliver valid combined estimates. The conclusions of Zucker et al. (2010) were that 'with few observations per patient and little information about within-patient variation, combined N-of-1 trials data may not support models that include complex variance structures.' If there are substantive concerns about the trials or the review, then it may be better not to undertake a meta-analysis at all. However, with sufficient information they can be used to estimate population effects and can provide enhanced estimates and inferences compared with standard clinical trials. Moreover, 'models with fixed treatment effects and common variances are robust and lead to conclusions that are similar to, though more precise, than single period or single crossover designs' (p. 1312).

In conclusion, the increasing interest in, and application of N-of-trials is clear. The systematic review reported by Gabler et al. (2011), based on 2154 participants in 108 studies published between 1985 and December 2010, found that N-of-1 trials were useful for increasing precision of estimates for a range of medical conditions. The User Guide for these trials authored by Kravitz et al. (2014a) and sponsored by the U.S. Agency for Healthcare Research and Quality provides further evidence. Systematic reviews and meta-analysis of these studies is the next logical step in evidence-based medicine. The basic guidelines and methods are described in this chapter, and elaborations are available (Duan et al. 2013; Schmid and Duan 2014; Zucker et al. 2010). It behoves the biostatisticians involved in these fields to keep developing, improving and applying them.

References

Dallery J, Cassidy RN, Raiff BR (2013) Single-case experimental designs to evaluate novel technology-based health interventions. J Med Internet Res 15(2), e22

Dallery J, Raiff BR (2014) Optimizing behavioral health interventions with single-case designs: from development to dissemination. Transl Behav Med 4:290–303

Duan N, Kravitz RL, Schmid CH (2013) Single-patient (N-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol 66:S21–S28

Gabler NB, Duan N, Vohra S, Kravitz RL (2011) N-of-1 trials in the medical literature: a systematic review. Med Care 49:761–768

Higgins J, Green SE (2008) Cochrane handbook for systematic reviews of interventions (eds) Wiley

Koricheva J, Gurevitch J, Mengersen K (eds) (2013) Hand-book of meta-analysis in ecology and evolution. Princeton University Press, Princeton

Kravitz RL, Duan N, The Decide Methods Centre N-of-1 Guidance Panel (Duan N, Eslick L, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S) (eds) (2014a) Design and implementation of N-of-1 trials: a user's guide. Agency for Healthcare Research and Quality, Rockville

Kravitz R, Duan N, Vohra S, Li J (2014b) The DEcIDE methods centre N-of-1 guidance panel introduction to N-of-1 trial: indications and barriers. In: Kravitz R, Duan N, The Decide Methods Centre N-of-1 Guidance Panel (Duan N, Eslick L, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S) (eds) Design and implementation of N-of-1 trials: a user's guide. Agency for Healthcare Re-search and Quality, Rockville

Lillie EO, Patay B, Diamant J, Isseli B, Topol EJ, Schoril NJ (2011) The N-of-1 clinical trial: the ultimate strategy for individualizing medicine? Pers Med 8(2):161–173

Schmid C, Duan N (2014) The DEcIDE methods centre N-of-1 guidance panel statistical design and analytic consideration for N-of-1 trials. In: Kravitz RL, Duan N, Eslick L, Gabler NB, Kaplan HC, Kravitz RL, Larson EB, Pace WD, Schmid CH, Sim I, Vohra S (eds) Design and implementation of N-of-1 trials: a user's guide. Agency for Healthcare Research and Quality, Rockville

Senior HE, Mitchell GK, Nikles J, Carmont SA, Schluter PJ, Currow DC, Vora R, Yelland MJ, Agar M, Good PD, Hardy JR (2013) Using aggregated single patient (N-of-1) trials to determine the effectiveness of psychostimulants to reduce fatigue in advanced cancer patients: a rationale and protocol. BMC Palliat Care 12:17

Yelland MJ, Poulos CJ, Pillans PI, Bashford GM, Nikles CJ, Sturtevant JM, Vine N, Del Mar CB, Schluter PJ, Tan M, Chan J, Mackenzie F, Brown R (2009) N-of-1 randomized trials to assess the efficacy of gabapentin for chronic neuropathic pain. Pain Med 10:754–761

Zucker DR, Schmid CH, Mcintosh MW, D'agostino RB, Selker HP, Lau J (1997) Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. J Clin Epidemiol 50:401–410

Zucker DR, Ruthazer R, Schmid CH, Feuer JM, Fischer PA, Kieval RI, Mogavero N, Rapoport RJ, Selker HP, Stotsky SA, Winston E, Goldenberg DL (2006) Lessons learned combining N-of-1 trials to assess fibromyalgia therapies. J Rheumatol 33:2069–2077

Zucker DR, Ruthazer R, Schmid CH (2010) Individual (N-of-1) trials can be combined to give population comparative treatment effect estimates: methodologic considerations. J Clin Epidemiol 63:1312–1323

Chapter 17
Where Are N-of-1 Trials Headed?

Jane Nikles

Abstract N-of-1 trials and review articles have recently been published in the areas of chronic pain, pediatrics, palliative care, complementary and alternative medicine, rare diseases, patient-centered care, the behavioral sciences and genomics. These are briefly reviewed and the current place of N-of-1 trials discussed. The chapter concludes with a vision for the future of N-of-1 trials.

Keywords N-of-1 trials • Chronic pain • Pediatrics • Palliative care • Complementary and alternative medicine • Rare diseases • Patient-centered care • The behavioral sciences • Genomics

Current and Recently Published N-of-1 Trials and Reviews

N-of-1 trials are slowly gaining traction as their usefulness in a variety of situations becomes more clearly recognized. As of mid 2015, there are 8 N-of-1 trials listed as currently or soon to be recruiting on clinical trials.gov, consisting of 6 currently recruiting and 2 not yet recruiting. Two of these are in cancer, three in rare diseases, and 2 in children. In chronic pain research, palliative care, pediatrics, complementary and alternative medicine, rare disease research and the behavioral sciences, the place of N-of-1 trials is being solidified and strengthened.

Chronic Pain

At a recent Initiative on Methods, Measurement, and Pain Assessment in Clinical Trials (IMMPACT) consensus meeting to discuss research designs for proof-of-concept chronic pain clinical trials, the advantages and disadvantages of more recent trial designs, including N-of-1 designs, enriched designs, adaptive designs, and

J. Nikles (✉)
School of Medicine, The University of Queensland, Ipswich, QLD, Australia
e-mail: uqjnikle@uq.edu.au

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6_17

sequential parallel comparison designs, were summarized (Gewandter et al. 2014). Limitations discussed in relation to N-of-1 trials include the potential lack of generalizability of a single patient's results, the need for short treatment and washout periods and longer overall duration of follow-up for each patient in the trial. However, it is possible to implement an N-of-1 trial in a multicenter community practice setting (Zucker et al. 1997), which could be valuable for studying available treatments in new indications. Treatment effect estimates obtained from combining the results of multiple N-of-1 trials may provide valuable early-stage Proof of Concept evidence.

Pediatrics

Pediatrics is ideally suited to N-of-1 trials, with small populations, frequent heterogeneity of response and the clear benefits for parents of having individual information about their child's response. ADHD is the most common condition studied in pediatric N-of-1 trials: there have been N-of-1 trials of stimulants for ADHD in a total of 193 children in 4 studies since 1996 (see Table 17.1). A further 138 children

Table 17.1 Drugs/conditions studied in pediatric N-of-1 trials to date

Author	Drug	Condition	Number of children
Nikles et al. 2006	CNS stimulants	ADHD	108 children
Duggan et al. 2000	CNS stimulants	ADHD	4 children
Faber et al. 2007	CNS stimulants	ADHD	31 children
Kent et al. 1999	methylphendiate	ADHD	50 children
Huber et al. 2007	Amitryptiline	Pain in juvenile idiopathic arthritis	6 children
Sung et al. 2007	Topical vitamin E	Prophylaxis for chemotherapy-induced oral mucositis	16 children; 45 post chemotherapy cycles were randomised to vitamin E (N = 22) or placebo (N = 23)
Nathan et al. 2006	Ondansetron plus metopimazine vs. ondansetron monotherapy	Children receiving highly emetogenic chemotherapy	12 children
Suri et al. 2004	RhDNase and hypertonic saline	Cystic fibrosis	48 children
Camfield et al. 1996	Melatonin	Intellectual deficits and fragmented sleep	6 children
Nikles et al. 2014	CNS Stimulants vs. placebo	Acquired Brain Injury	50 children: 10 published in Nikles et al. 2014 and 40 unpublished (unpublished data, Nikles et al.)

have undertaken N-of-1 trials in a variety of conditions, making a total of 331 children undergoing N-of-1 trials since 1996. The various drugs/conditions studied are listed in table 17.1.

There have been four reviews about the use of N-of-1 trials in children for

- Complementary and alternative medicines in cancer Sung and Feldman (2006).
- Human deoxyribonuclease (rhDNase) in the management of cystic fibrosis (Suri 2005)
- Montelukast in pediatric asthma (Bush 2014)
- Psychopharmacological studies (Greenhill et al. 2003).

More recently, attention has turned to pediatric analgesic trials. The standard parallel-placebo analgesic trial design commonly used for adults has scientific, ethical and practical difficulties in pediatrics, due to the likelihood of subjects experiencing pain for extended periods of time. Participants in a FDA sponsored scientific workshop developed consensus on aspects of pediatric analgesic clinical trial design. The consensus was that small sample designs, including cross-over trials and N-of-1 trials, should be considered for particular pediatric chronic pain conditions and for studies of pain and irritability in pediatric palliative care (Berde et al. 2012). One option is to compare best analgesia vs best analgesia plus test treatment, which removes the ethical problem of placebos for pain trials.

Palliative Care

N-of-1 trials are a new methodology well suited to meet some of the challenges of conducting trials in a palliative care (PC) setting (Davis and Mitchell 2012). The need to improve the evidence base on which PC is based is widely acknowledged (Hermet et al. 2002), especially as many of the common practices and interventions used routinely are based on anecdote or expert opinion alone. RCTs are considered by many to be the gold standard for evidence in clinical medicine. However, many RCTs fail in palliative populations (e.g. Cook et al. 2002) because it is too difficult to recruit and retain enough people to achieve the predicted sample size without extraordinary amounts of effort, organization and funding. Multi-site support is needed, as patients want to participate. N-of-1 trials are an alternative means of conducting trials in these patients. There is also the issue of how to manage missing data, where a large proportion of the patient population is likely to die. In normal intention to treat trials, these are considered as treatment failures, and this is not the case in PC (Currow et al. 2012). Utilizing N-of-1 trials and including all completed cycles in the final analysis overcomes this problem.

As described in Chap. 16, it is possible to combine the results of many N-of-1 trials to determine what the effect of a therapy was for a population (Zucker et al. 1997). N-of-1 trials can gather evidence of similar strength to RCTs in PC, but require less than half the number of subjects. This allows more rapid accumulation of strong evidence on treatment effects in patients with advanced life-limiting illness previously very difficult to gather. For suitable clinical questions, N-of-1 trials will enable high quality evidence to be gathered much more effectively, accelerate the rate of accumulation of high-grade evidence and have an important effect on the quality and effectiveness of care offered to this very disadvantaged group (Mitchell et al. unpublished data).

Complementary and Alternative Medicine

N-of-1 trials have been used to test valerian for insomnia (Coxeter et al. 2003). Recently several articles have been published by Chinese groups about using N-of-1 trials for Traditional Chinese Medicine (TCM) (Li et al. 2013; Huang et al. 2014). One example is Liuwei Dihuang- Decoction for Kidney-Yin Deficiency Syndrome (Yuhong et al. 2013). N-of-1 trials are uniquely suited to the individualized nature of TCM, though limited information about the half-lives of some of these medicines make estimating the length of each treatment period and washout periods, and therefore trial length, difficult. They may be suitable for trials of acupuncture using a sham needle (Lee et al. 2012).

Rare Diseases

Rare diseases may be difficult to study through conventional research methods, because of the small numbers of patients, but are well suited to study through N-of-1 trials (Gupta et al. 2011). N-of-1 trials could be particularly valuable for rare diseases when prospectively planned across several patients and analyzed using Bayesian techniques; a population effect can then be estimated that will be of value to Health Technology Assessment (Facey et al. 2014). Multi-site trials and storage of patient data that could be combined with patient data from future trials is important in this area.

Patient-Centered Care

N-of-1 trials are a patient-centered intervention that may improves medication management in suitable chronic diseases. We conducted the first study examining patient perspectives of N-of-1 trials (Nikles et al. 2005).. Patients were generally very satisfied with the N-of-1 trial process. Their participation led to increased knowledge, awareness and understanding of their condition, their bodies' response to it, and its management. Some of this arose specifically from use of daily symptom diaries. N-of-1 trials led to a sense of empowerment and control as well as improved individually-focused care. N-of-1 trials appeared to empower these patients as a result of both collecting information about their responses to different treatment options, and participating actively in subsequent therapeutic decisions.

Taragin et al. (2013) compared parents' attitudes toward methylphenidate treatment in children with Attention Deficit Hyperactivity Disorder employing two approaches: (1) a 2-week double-blind placebo-drug trial (N-of-1 trial), and (2) a traditional prescription approach. While initial attitudes were similar, a significantly more favorable attitude following an N-of-1 trial and throughout the follow-up of this group was found. Adherence was significantly correlated with attitude score in the N-of-1 group only. An individual N-of-1 trial with methylphenidate appeared to positively affect parents' attitudes toward drug treatment and also adherence with this treatment.

N-of-1 trials may soon emerge as an important part of the methodological armamentarium for comparative effectiveness research and patient-centered outcomes research. By permitting direct estimation of individual treatment effects, they allow finely tuned individualized care, and can enhance therapeutic precision, improve patient outcomes, and reduce costs (Duan et al. 2013).

Davidson et al. propose increased use of N-of-1 trials in situations where treatment response is heterogeneous – as is the case for most psychological and behavioral treatments (Davidson et al. 2014). Davidson and team have recently been funded for a project called Engaging Stakeholders in Building Patient-Centered, N-of-1 Randomized and Other Controlled Trial Methods. The objectives of the study are to identify a promising set of medical conditions and methodologies for N-of-1 RCTs, create educational materials to inform patients of the pros and cons of these trials, and determine which directions these methods should take to be most useful to patients. The study will engage patients and other key stakeholders (clinicians, researchers, statisticians, pharmacists, and ethicists) to prospectively shape the research and methods agenda of an N-of-1 RCT approach. The results of this research will allow comparison of the effect of conducting N-of-1 RCTs versus usual care on patient-chosen outcomes such as symptoms, disease control, and satisfaction with care.

Behavioral Sciences

Recent publication of several articles outlining aspects of the conduct, quality assessment and interpretation and the process of developing reporting guidelines for the conduct of SCEDs (Single Case Experimental Designs) in the behavioral sciences reflect the resurgence of interest in this design, not only in medicine but the behavioral sciences (Tate et al. 2008, 2013, 2014; Evans et al. 2014).

Genomics

Medicine has moved towards personalized or "precision medicine"; there is an upsurge in pharmacogenomics studies. Prediction of response/non-response and adverse events requiring cessation or switching of therapy for important drugs would be of enormous health and economic benefit. Usually large numbers of patients require genetic testing to unravel gene sequences. Using genomes to predict response has been carried out for various drugs e.g. warfarin, azathioprine, some cancer drugs, clopidogrel. Applying N-of-1 trials to pharmacogenomics, which would significantly reduce sample size required, has not yet been done, though suggested in a number of articles (Lillie et al. 2011).

We conclude with a quote from Kaput and Morine 2012 who are developing N-of-1 nutrigenomic research:

> High throughput metabolomics, proteomic and genomic technologies provide 21st century data that humans cannot be randomized into groups: individuals are genetically and biochemically distinct. Gene–environment interactions caused by unique dietary and lifestyle factors contribute to heterogeneity in physiologies observed in human studies. The risk factors determined for populations cannot be applied to the individual. Developing individual risk or benefit factors in light of the genetic diversity of human populations, the complexity of foods, culture and lifestyle, and the variety of metabolic processes that lead to health or disease are significant challenges for personalizing advice for healthy or medical treatments for individuals with chronic disease (Kaput and Morine 2012)

So Where Are N-of-1 Trials Going?

Vision

Imagine this*……a patient attends their doctor with any chronic disease, e.g. osteoarthritis. Before prescribing a medication, the doctor writes a "prescription" for an N-of-1 trial, a test to see whether the medication works for the patient's pain. The trial is set up on a mobile phone app allowing customized design of the trial. After taking medication and placebo in blinded random order and keeping track of pain/symptoms via the mobile phone app, the patient and their doctor receive a report about whether the drug works for their pain and whether it has side effects. N-of-1 trials are widely known, and standard practice in clinical situations where there is uncertainty about the effectiveness of a drug, there is uncertainty about the dose that will be effective, the drug is expensive or it has important side effects. Patients initiate discussion with their doctor about using N-of-1 trials to answer specific questions about their health. In rare conditions, conditions where recruitment is difficult or populations are small, N-of-1 trials, the highest level of evidence, are commonly used to assess effectiveness of drugs where the drug to be tested is suitable. Pharmaceutical funders such as health insurers and state government health services use N-of-1 trials to decide whether a patient responds and therefore should have the cost of the drug reimbursed. A central coordinating unit runs N-of-1 trials all over the country by post and telephone, working closely with a manufacturing pharmacy to supply medications. N-of-1 trials are used in many countries, with a national coordinating center in each country. A worldwide database stores the design and results of each N-of-1 trial for aggregation with other similar trials to facilitate the application of sophisticated statistical methods to analyse the trials.*

Conclusion

N-of-1 trials are becoming more widely used, and their application in certain suitable areas such as pediatrics and chronic pain is growing. The confluence of genomics, the upswing in personalized medicine and the widespread popularity of wireless devices make a promising platform for N-of-1 trials to find their true niche.

References

Berde CB, Walco GA, Krane EJ, Anand KJ, Aranda JV, Craig KD, Dampier CD, Finkel JC, Grabois M, Johnston C, Lantos J, Lebel A, Maxwell LG, McGrath P, Oberlander TF, Schanberg LE, Stevens B, Taddio A, Von Baeyer CL, Yaster M, Zempsky WT (2012) Pediatric analgesic clinical trial designs, measures, and extrapolation: report of an FDA scientific workshop. Pediatrics 129(2):354–364. doi:10.1542/peds.2010-3591, Epub 2012 Jan 16

Bush A. Montelukast in paediatric asthma: where we are now and what still needs to be done? 7 2014 Dec 11. pii: S1526-0542(14)00131-6. doi: 10.1016/j.prrv.2014.10.007

Camfield P, Gordon K, Dooley J, Camfield C (1996) Melatonin appears ineffective in children with intellectual deficits and fragmented sleep: six "N of 1" trials. J Child Neurol 11(4):341–343

Cook AM, Finlay IG, Butler-Keating RJ (2002) Recruiting into palliative care trials: lessons learnt from a feasibility study. Palliat Med 16:163–165

Coxeter P, Schluter P, Eastwood H, Nikles J, Glasziou P (2003) Valerian does not reduce symptoms for patients with chronic insomnia in general practice. Complement Ther Med 11(4):215–222

Currow DC, Plummer JL, Kutner JS, Samsa GP, Abernethy AP (2012) Analyzing phase III studies in hospice/palliative care. A solution that sits between intention-to-treat and per protocol analyses: the palliative-modified ITT analysis. J Pain Symptom Manag 44(4):595–603

Davidson KW, Peacock J, Kronish IM, Edmondson D (2014) Personalizing behavioral interventions through single-patient (N-of-1) trials. Soc Personal Psychol Compass 8(8):408–421

Davis MP, Mitchell GK (2012) Topics in research: structuring studies in palliative care. Curr Opin Support Palliat Care 6(4):483–489. doi:10.1097/SPC.0b013e32835843d7

Duan N, Kravitz RL, Schmid CH (2013) Single-patient (N-of-1) trials: a pragmatic clinical decision methodology for patient-centered comparative effectiveness research. J Clin Epidemiol 66(8 Suppl):S21–S28. doi:10.1016/j.jclinepi.2013.04.006

Duggan CM, Mitchell G, Nikles CJ, Glasziou PP, Del Mar CB, Clavarino A (2000) Managing ADHD in general practice. N of 1 trials can help! Aust Fam Physician 29(12):1205–1209

Evans JJ, Gast DL, Perdices M, Manolov R (2014) Single case experimental designs: introduction to a special issue of neuropsychological rehabilitation. Neuropsychol Rehabil 24(3–4):305–314. doi:10.1080/09602011.2014.903198, Epub 2014 Apr 25

Faber A, Keizer RJ, van den Berg PB, de Jong-van den Berg LT, Tobi H (2007) Use of double-blind placebo-controlled N-of-1 trials among stimulant-treated youths in the Netherlands: a descriptive study. Eur J Clin Pharmacol 63(1):57–63, Epub 2006 Nov 18

Facey K, Granados A, Guyatt G, Kent A, Shah N, van der Wilt GJ, Wong-Rieger D (2014) Generating health technology assessment evidence for rare diseases. Int J Technol Assess Health Care 19:1–7

Gewandter JS, Dworkin RH, Turk DC, McDermott MP, Baron R, Gastonguay MR, Gilron I, Katz NP, Mehta C, Raja SN, Senn S, Taylor C, Cowan P, Desjardins P, Dimitrova R, Dionne R, Farrar JT, Hewitt DJ, Iyengar S, Jay GW, Kalso E, Kerns RD, Leff R, Leong M, Petersen KL, Ravina BM, Rauschkolb C, Rice AS, Rowbotham MC, Sampaio C, Sindrup SH, Stauffer JW, Steigerwald I, Stewart J, Tobias J, Treede RD, Wallace M, White RE (2014) Research designs for proof-of-concept chronic pain clinical trials: IMMPACT recommendations. Pain 155(9):1683–1695. doi:10.1016/j.pain.2014.05.025, Epub 2014 May 24

Greenhill LL, Jensen PS, Abikoff H, Blumer JL, Deveaugh-Geiss J, Fisher C, Hoagwood K, Kratochvil CJ, Lahey BB, Laughren T, Leckman J, Petti TA, Pope K, Shaffer D, Vitiello B, Zeanah C (2003) Developing strategies for psychopharmacological studies in preschool children. J Am Acad Child Adolesc Psychiatry 42(4):406–414, Review

Gupta S, Faughnan ME, Tomlinson GA, Bayoumi AM (2011) A framework for applying unfamiliar trial designs in studies of rare diseases. J Clin Epidemiol 64(10):1085–1094. doi:10.1016/j.jclinepi.2010.12.019, Epub 2011 May 6

Hermet R, Burucia B, Sentilles-Monkam A (2002) The need for evidence-based proof in palliative care. Eur J Pall Care 9:104–107

Huang H, Yang P, Xue J, Tang J, Ding L, Ma Y, Wang J, Guyatt GH, Vanniyasingam T, Zhang Y (2014) Evaluating the individualized treatment of traditional Chinese medicine: a pilot study of N-of-1 trials. Evid Based Complement Alternat Med 2014:148730. doi:10.1155/2014/148730, Epub 2014 Nov 11

Huber AM, Tomlinson GA, Koren G, Feldman BM (2007) Amitriptyline to relieve pain in juvenile idiopathic arthritis: a pilot study using Bayesian metaanalysis of multiple N-of-1 clinical trials. J Rheumatol 34(5):1125–1132, Epub 2007 Apr 15

Kaput J, Morine M (2012) Discovery-based nutritional systems biology: developing N-of-1 nutrigenomic research. Int J Vitam Nutr Res 82(5):333–341. doi:10.1024/0300-9831/a000128

Kent MA, Camfield CS, Camfield PR (1999) Double-blind methylphenidate trials: practical, useful, and highly endorsed by families. Arch Pediatr Adolesc Med 153(12):1292–1296

Lee S, Lim N, Choi SM, Kim S (2012) Validation study of Kim's sham needle by measuring facial temperature: an N-of-1 randomized double-blind placebo-controlled clinical trial. Evid Based Complement Alternat Med 2012:507937. doi:10.1155/2012/507937, Epub 2012 Mar 6

Li J, Tian J, Ma B, Yang K (2013) N-of-1 trials in china. Complement Ther Med 21(3):190–194. doi:10.1016/j.ctim.2013.01.003, Epub 2013 Feb 18

Lillie EO, Patay B, Diamant J, Issell B, Topol EJ, Schork NJ (2011) The N-of-1 clinical trial: the ultimate strategy for individualizing medicine? Per Med 8(2):161–173

Nathan PC, Tomlinson G, Dupuis LL, Greenberg ML, Ota S, Bartels U, Feldman BM (2006) A pilot study of ondansetron plus metopimazine vs. Ondansetron Monotherapy in children receiving highly emetogenic chemotherapy: a Bayesian randomized serial N-of-1 trials design. Support Care Cancer 14(3):268–276, Epub 2005 Jul 29

Nikles CJ, Clavarino AM, Del Mar CB (2005) Using N-of-1 trials as a clinical tool to improve prescribing. Br J Gen Pract 55(512):175–180

Nikles CJ, Mitchell GK, Del Mar CB, Clavarino AM, McNairn N (2006) An n-of-1 trial service in clinical practice: testing the effectiveness of stimulants for attention-deficit/hyperactivity disorder. Pediatrics 117(6):2040–2046

Nikles CJ, McKinlay L, Mitchell GK, Carmont SA, Senior HE, Waugh MC, Epps A, Schluter PJ, Lloyd OT (2014) Aggregated n-of-1 trials of central nervous system stimulants versus placebo for paediatric traumatic brain injury–a pilot study. Trials 15:54. doi:10.1186/1745-6215-15-54

Sung L, Feldman BM (2006) N-of-1 trials: innovative methods to evaluate complementary and alternative medicines in pediatric cancer. J Pediatr Hematol Oncol 28(4):263–266

Sung L, Tomlinson GA, Greenberg ML, Koren G, Judd P, Ota S, Feldman BM (2007) Serial controlled N-of-1 trials of topical vitamin E as prophylaxis for chemotherapy-induced oral mucositis in paediatric patients. Eur J Cancer 43(8):1269–1275, Epub 2007 Mar 23

Suri R (2005) The use of human deoxyribonuclease (rhDNase) in the management of cystic fibrosis. BioDrugs 19(3):135–144

Suri R, Metcalfe C, Wallis C, Bush A (2004) Predicting response to rhDNase and hypertonic saline in children with cystic fibrosis. Pediatr Pulmonol 37(4):305–310

Taragin D, Berman S, Zelnik N, Karni A, Tirosh E (2013) Parents' attitudes toward methylphenidate using n-of-1 trial: a pilot study. Atten Defic Hyperact Disord 5(2):105–109. doi:10.1007/s12402-012-0099-x. Epub 2012 Dec 16

Tate RL, McDonald S, Perdices M, Togher L, Schultz R, Savage S (2008) Rating the methodological quality of single-subject designs and n-of-1 trials: introducing the single-case experimental design (SCED) scale. Neuropsychol Rehabil 18(4):385–401. doi:10.1080/09602010802009201

Tate RL, Perdices M, Rosenkoetter U, Wakim D, Godbee K, Togher L, McDonald S (2013) Revision of a method quality rating scale for single-case experimental designs and n-of-1 trials: the 15-item risk of bias in N-of-1 trials (RoBiNT) scale. Neuropsychol Rehabil 23(5):619–638. doi:10.1080/09602011.2013.824383, Epub 2013 Sep 9

Tate RL, Perdices M, McDonald S, Togher L, Rosenkoetter U (2014) The design, conduct and report of single-case research: resources to improve the quality of the neurorehabilitation literature. Neuropsychol Rehabil 24(3–4):315–331. doi:10.1080/09602011.2013.875043, Epub 2014 Apr 7

Yuhong H, Qian L, Yu L, Yingqiang Z, Yanfen L, Shujing Y, Shufang Q, Lanjun S, Shuxuan Z, Baohe W (2013) An n-of-1 trial service in clinical practice: testing the effectiveness of Liuwei Dihuang decoction for kidney-Yin defisciency syndrome. Evid Based Complement Alternat Med 2013:827915. doi:10.1155/2013/827915, Epub 2013 Sep 23

Zucker DR, Schmid CH, McIntosh MW, D'Agostino RB, Selker HP, Lau J (1997) Combining single patient (N-of-1) trials to estimate population treatment effects and to evaluate individual patient responses to treatment. J Clin Epidemiol 50(4):401–410

Index

J. Nikles, G. Mitchell (eds.), *The Essential Guide to N-of-1 Trials in Health*,
DOI 10.1007/978-94-017-7200-6