ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Statistical power in two-level hierarchical linear models with arbitrary number of factor levels

Yongyun Shin [a,*], Jennifer Elston Lafata [b], Yu Cao [c]

[a] Department of Biostatistics, Virginia Commonwealth University, P.O. Box 980032, 830 East Main Street, Richmond, VA 23298-0032, United States
[b] University of North Carolina at Chapel Hill, United States
[c] Virginia Commonwealth University, United States

ARTICLE INFO

Article history:
Received 2 January 2017
Received in revised form 5 September 2017
Accepted 5 September 2017
Available online xxxx

Keywords:
Cluster randomized trial
Multisite randomized trial
Split-plot randomized trial
Interaction effects
Omnibus test
Sample sizes

ABSTRACT

As the US health care system undergoes unprecedented changes, the need for adequately powered studies to understand the multiple levels of main and interaction factors that influence patient and other care outcomes in hierarchical settings has taken center stage. We consider two-level models where n lower-level units are nested within each of J higher-level clusters (e.g. patients within practices and practices within networks) and where two factors may have arbitrary a and b factor levels, respectively. Both factors may represent $a \times b$ treatment combinations, or one of them may be a pretreatment covariate. Consideration of both factors at the same higher or lower hierarchical level, or one factor per hierarchical level yields a cluster (C), multisite (M) or split-plot randomized design (S). We express statistical power to detect main, interaction, or any treatment effects as a function of sample sizes (n, J), a and b factor levels, intraclass correlation ρ and effect sizes δ given each design $d \in \{C, M, S\}$. The power function given a, b, ρ, δ and d determines adequate sample sizes to achieve a minimum power requirement. Next, we compare the impact of the designs on power to facilitate selection of optimal design and sample sizes in a way that minimizes the total cost given budget and logistic constraints. Our approach enables accurate and conservative power computation with a priori knowledge of only three effect size differences regardless of how large $a \times b$ is, simplifying previously available computation methods for health services and other researches.

1. Introduction

The Affordable Care Act (ACA) and other pressures within the American health care market have created unprecedented incentives for physician practices and other health care providers to devise new ways of delivering care to patients. With these care innovations and the desire to understand their impact has come an increasing awareness that barriers to effectiveness can arise at multiple hierarchical levels including the patient, provider-team, organization, and local market or policy levels (Ferlie and Shortell, 2001; Taplin et al., 2012; Nutting et al., 2011; Miller et al., 2010). The inherently hierarchical structure of the health care system demands multilevel research methods that enable an understanding of how the multiple hierarchical contexts affect the impact of new care innovations as well as an understanding of the interplay between characteristics at the different hierarchical levels. The latter is particularly important to reducing known disparities in care outcomes, as it is the latter that enables an understanding of how different care innovations or organizational contexts

* Corresponding author.
E-mail address: yshin@vcu.edu (Y. Shin).

https://doi.org/10.1016/j.jspi.2017.09.004
0378-3758/

Please cite this article in press as: Shin Y., et al., Statistical power in two-level hierarchical linear models with arbitrary number of factor levels. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.004.

may impact outcomes for vulnerable or otherwise important population subgroups (e.g., those of lower socio-economic status, from a minority race, or with multiple chronic conditions).

Furthermore, the impact of care innovations is highly dependent on the ability to successfully implement them across diverse settings and patient populations (Damschroder et al., 2009; Glasgow et al., 2001). As such, it is often important to consider not just whether or not a population was exposed to a given care innovation (i.e., treatment), but the extent to which they were exposed (i.e., treatment intensities or factor levels) (McHugh et al., 2015). Therefore, it is important for statistical methods, including power, to consider the possibility of not just the presence or absence of a given care innovation but the multiple relevant intensities of that innovation.

Despite the advances in statistical methods for analyzing hierarchical or multilevel data (Raudenbush and Bryk, 2002; Goldstein, 2003; Snijders and Bosker, 2012), recent multilevel power analyses have focused mainly on a single binary treatment factor in either multisite or cluster randomized designs (Donner and Klar, 1996; Raudenbush, 1997; Raudenbush and Liu, 2000, 2001; Raudenbush et al., 2007; Hedeker et al., 1999; Moerbeek, 2008; Schochet, 2008; Heo and Leon, 2008; Konstantopoulos, 2009; Hedges, 2011; Usami, 2014a, b; Cunningham and Johnson, 2016), thereby limiting their applicability in the context of studying complex health care innovations. It is not surprising, therefore, that recent evaluations of new care models and other system redesign strategies have often been hindered by inadequate statistical power (Jackson et al., 2013; Peikes et al., 2011).

Usami (2011) considered a multisite design where two factors may represent the arbitrary number of factor levels and where power to detect the main or interaction effects is based on multiparameter contrasts using Wald statistics (Raudenbush and Bryk, 2002). Although useful for large samples, Usami's formula may yield inaccurate power for small samples, and depends on sample sizes, effect sizes and variance components. As we will show, the exact power will, in addition, depend on the number of factor levels and vary widely across designs.

We derive power to detect main, interaction, or any (i.e. at least one) treatment effects of factors A and B in a two-level hierarchical linear model (HLM), also known as a multilevel or mixed effects model, where the factors represent arbitrary a and b factor levels or treatment intensities, respectively (Raudenbush and Bryk, 2002; Goldstein, 2003; Snijders and Bosker, 2012). One of the factors may be a pretreatment covariate such as the size of a practice, availability of electronic health records, and the patient's socio-economic status or race/ethnicity. Randomizing both factors at the same higher or lower hierarchical level, or one factor per hierarchical level yields a cluster (C), multisite (M) or split-plot randomized design (S) (Montgomery, 2005). Consequently, statistical power depends on the type of design $d \in \{C, M, S\}$. We express power as a function of sample sizes, a and b factor levels, intraclass correlation ρ, effect sizes δ, and design d, and compare the impact of the determinants on power.

Given the determinants without constraints, a multisite design produces higher power than other designs. It is important, however, to learn the impact of designs on power as sample sizes are typically constrained due to logistic and/or cost constraints, in particular, when $a \times b$ treatment combinations are large. For example, solo or small physician practices can provide a limited number of patients for any given study, and access to physician practices or other higher level clusters may be even more limited. Such practical limitations make comparison of power across the designs relevant and useful. Given a, b, ρ, δ and constraints, the power function determines optimal design and sample sizes in a way that achieve a minimum power requirement and, also, in a way that minimize the research costs. In particular, we derive a single uniform closed-form formula for power to detect main, interaction or any treatment effects given each design. The resulting formula facilitates not only computation of power to detect any of the effects given any of the designs, but also comparison of power across the designs given logistic and cost constraints. Throughout this paper, we assume perfect compliance to the treatment assignment (Angrist et al., 1996).

Illustrative of the methodological challenges faced was recent request by the Agency for Health Care Research and Quality (AHRQ) for applications to evaluate programs designed to accelerate the dissemination and implementation of results from patient-centered, evidence-based cardio-vascular care outcomes research findings into primary care (http://grants.nih.gov/grants/guide/rfa-files/RFA-HS-14-009.html). Under this Initiative, eight "Health Cooperatives (Coops)" or "networks" each bringing with them some 250 primary care practices, are used to test how externally provided quality improvement (QI) support (i.e., treatment) can accelerate the dissemination and implementation of patient-centered, evidence-based care. By design (http://grants.nih.gov/grants/guide/rfa-files/RFA-HS-14-008.html), the Coops differ in their QI support approaches. That diversity and vast number of practices created a unique opportunity for an external, cross-cutting evaluation to provide new insights into how different QI support approaches, and primary care practices' scope and intensity of engagement with those supports, could impact outcomes among diverse practices and patients. Central to a comprehensive evaluation is an understanding not only of the Initiative's overall effectiveness, but also the comparative effectiveness of different QI support strategies for different practices, recognizing that practices will differ in their intensity of engagement with the Initiative. Thus, in addition to using a simple binary variable reflective of exposure to the Initiative for a given practice in a given time period (i.e., overall treatment vs. comparison group contrasts), it would be advantageous to consider both types of QI support approaches available (i.e., "treatment") and each practice's level of exposure to them (i.e., "treatment intensity" such as how frequently they participated in each activity and their level of commitment to the activity). Additionally, when planning such an evaluation, it is important to consider the cost implications of the number of Coops, practices and patients. We revisit design of this evaluation in Section 6.

Because our power computation is not approximate, but exact, it is useful for both large and small sample sizes (Raudenbush and Liu, 2001). Explicit dependence of the power function on the arbitrary a and b treatment intensities and designs

Please cite this article in press as: Shin Y., et al., Statistical power in two-level hierarchical linear models with arbitrary number of factor levels. J. Statist. Plann. Inference (2017), https://doi.org/10.1016/j.jspi.2017.09.004.