Available online at www.sciencedirect.com

ScienceDirect

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

Power and Sample Size Calculations in Clinical Trials with Patient-Reported Outcomes under Equal and Unequal Group Sizes Based on Graded Response Model: A Simulation Study

*Marziyeh Doostfatemeh, Seyyed Mohammad Taghi Ayatollah**, *Peyman Jafari*

Department of Biostatistics, Shiraz University of Medical Sciences, Shiraz, Iran

ABSTRACT

Objectives: To provide a valid sample size strategy based on simulation and to evaluate the statistical power in clinical trials with patient-reported outcomes (PROs) based on a polytomous item response theory model—the graded response model (GRM)—and to compare this framework with the classical test theory (CTT) approach. **Methods:** One thousand randomized clinical trials were simulated using PRO based on the GRM and under various combinations of the number of patients in each arm, the group allocation ratio, the number of items and categories, and group effects. The power and sample size estimated in the simulations were then compared with those computed using the CTT framework. **Results:** The results indicated that the impact of the most influential factors, including the number of patients, group allocation ratio, group effects, and the number of categories, on the power and sample size of the GRM-based and CTT-based approaches was similar. Nevertheless, the strong impact of the number of items on these issues distinguished the two approaches. **Conclusions:** It is crucial to use an adapted sample size formula in a GRM-based analysis because the classical formula designed for the CTT-based approach does not consider the impact of the number of items, which could result in an inadequately sized study and a decrease in power. Thus, when clinicians design a randomized clinical trial with polytomous PRO endpoints using classical sample size formula as the base, they should be aware of the possibility of making an incorrect clinical decision.

Keywords: graded response model, patient-reported outcome, power, sample size.

Introduction

Patient-reported outcomes (PROs) are now recognized as important endpoints in randomized clinical trials (RCTs) [1]. The initial focus of an RCT has evolved from traditional efficacy and safety evaluation of new treatments [1,2]; nevertheless, information about outcomes that can be reported only by patients is also important to measure, both in RCTs and in routine clinical care [3–6]. In this sense, it is recommended to record PROs such as health-related quality of life in all RCTs as primary or key secondary endpoints [5–9].

Two types of framework are used for analyzing PROs: the classical test theory (CTT) and methods derived from the item response theory (IRT) [10]. In the CTT framework, PROs are evaluated using self-assessment questionnaires from which the item responses provided by patients are combined to give observed scores. In this approach, the observed score is a combination of the underlying "true score" of a PRO instrument and random measurement error. The observed score is a good estimation of the true score; PROs, however, are latent in nature and cannot be observed directly. Therefore, the use of a statistical model related to observed variables has been considered to be faulty and must be analyzed using an appropriate modeling strategy [11]. The IRT framework was developed to rectify dissatisfaction with the earlier methods [12]. This method is based on a response model linking the item responses to a latent parameter. The latent trait in IRT-based models is equivalent to the true score in CTT-based analyses [13]. The major difference between observed and latent variable models (IRT vs. CTT) is in accounting for or ignoring measurement error. The topic of measurement error correction in observed variables has received much attention elsewhere [14,15].

The main advantage of using IRT over CTT for PRO data from RCTs is that IRT-based models provide a powerful framework to construct and reduce PRO instruments and analyze data in a more accurate and efficient manner [16]. The graded response model (GRM) [17] is one of the most well-known polytomous IRT models. It is widely used in health research and has shown interesting psychometric properties, particularly sample-independent features. This means that, unlike the observed

Conflicts of interest: This article was extracted from M. Doostfatemeh's doctoral thesis.

* *Address correspondence to*: Seyyed Mohammad Taghi Ayatollahi, Department of Biostatistics, Shiraz University of Medical Sciences, Emam Hossein Square, Shiraz, Iran.

E-mail: ayatolahim@sums.ac.ir.

1098-3015$36.00 – see front matter Copyright © 2016, International Society for Pharmacoeconomics and Outcomes Research (ISPOR). Published by Elsevier Inc.

http://dx.doi.org/10.1016/j.jval.2016.03.1857

score, the estimated latent trait in the GRM is not seriously affected by the population and instruments. This property allows comparison of PROs from different populations and instruments. This is exemplified by computerized adaptive testing, in which each patient receives a different computer-administered questionnaire and the items offered to each patient depend on the responses given to the previous items [11,18]. In other words, in a particular patient population, it is feasible to obtain comparable estimates of PROs (e.g., health-related quality of life) using the most informative set of items for each patient.

A recent literature review uncovered a lack of psychometric sample size guidelines and requirements for the development of PROs for use in RCTs [19]. During an RCT, ethical considerations require that as few patients as possible are exposed to the risk of a new therapy; nevertheless, for PRO data from an RCT to have value, it is essential to ensure that the study size is large enough to accommodate the primary endpoint and assessment of PROs [19,20]. Hence, the use of an appropriate sample size calculation strategy is necessary to avoid wasting clinical resources and exposing patients to inappropriate medical decisions. It is strongly recommended that methods used for statistical analysis and sample size design should be based on similar methodological fields. Accordingly, although GRM-based analysis is more appropriate for fitting data from a PRO instrument, this framework must be considered early in the planning phases of the study, especially at the sample size calculation stage.

Previous IRT-based researches have shown that using the CTT-based strategy for observed variables underestimates sample size and significantly decreases power [21–24]. These studies focused on two-group cross-sectional designs and a binary-response model for analysis, although most PRO instruments used in RCTs consist of polytomous-response items. The present study, being based on simulation, therefore, provides the necessary advancements to adapt earlier methodologies to PRO data derived from clinical trials. In particular, it evaluates the power of GRM in comparison with a CTT-based approach in an RCT with PROs as the primary endpoint and provides guidelines to determine the number of patients required in each arm of an RCT. The study also considers the effect of group inequality on power and sample size calculations; these frequently occur in RCTs and could be helpful from ethical and economic aspects [25,26].

Methods

IRT Models

IRT models are mixed models in which the latent trait is randomly distributed and varies across patients [13]. These models are commonly used for PRO data and comprise two item parameters (discrimination and threshold) and one patient parameter (latent trait) [12]. The simple Rasch model is often used for binary-response questionnaires, and polytomous IRT models, including the rating scale model, partial credit model, generalized partial credit model (GPCM), and GRM, are used for multiple-response questionnaires. The rating scale model and the partial credit model are generally more restrictive than the GPCM and the GRM; for example, no discrimination parameter is available in these models and all items must have the same number of options [27]. The GPCM and the GRM cover a broader area for PRO data from RCTs and provide a more accurate description of such data. Both the GRM and the GPCM have their own functions to describe the probability of selecting a response category; therefore, model parameters cannot be compared directly between models [12]. Because the GRM is the polytomous IRT model most widely used in health research, it was used in the present study to model patient response probability to a latent trait.

Graded Response Model

Assume an RCT in which N patients have answered a questionnaire containing I polytomous items. Let X_{ik} be a random variable representing the response from patient k to item i and θ_k represent the latent trait for this patient. In the GRM, the probability of patient k responding at category j or higher of item i is calculated as follows:

$$P_{ijk}\left(X_{ik} \geq j \mid \theta_k, a_i, b_{ij}\right) = \frac{\exp\left[a_i\left(\theta_k - b_{ij}\right)\right]}{1+\exp\left[a_i\left(\theta_k - b_{ij}\right)\right]}, \tag{1}$$

where a_i is the item discrimination parameter and b_{ij} is the between-category threshold parameter of item i [13]. A large value for a_i indicates that the item is good at discriminating between levels of a trait. A large positive value for b_{ij} demonstrates a difficult item; that is, patients with a high level of a trait are more likely to give responses that fall into the higher categories. As for the IRT models, the GRM assumes that the items are unidimensional and the responses to them are independent of each other (local independence).

In the GRM, the patient and item parameters are typically estimated using a two-stage process. First, the item parameters are estimated by assuming that θ_k follows a Gaussian distribution and integrates them out of likelihood. Second, the maximum-likelihood estimates of θ_k are obtained using the item parameters estimated in the previous stage [28]. It is worth mentioning that accurate estimates of item parameters in GRM can be achieved only from large samples of patients (two to several hundreds) [11]. Because this figure is hardly ever attained in RCTs, it will not often be feasible to estimate the item parameters. Accordingly, the present study was designed to estimate the person parameters by assuming a fixed set of item parameters, such as in computerized adaptive testing [18], having item parameters, and figuring out the best items to administer [29,30]. That is, items under examination are from a calibrated item bank [31].

Sample Size Calculation: GRM-Based Framework

Suppose we plan to design an RCT using a given dimension of a PRO (e.g., physical functioning on the PedsQL questionnaire [32]) as the primary endpoint using the GRM. Let N_1 and N_2 be the sample sizes expected in each arm of the RCT and $N = N_1 + N_2$. Let θ be a latent trait with normal distributions $N\left(\mu_1 = -\frac{N_2}{N}d, \sigma^2\right)$ and $N\left(\mu_2 = -\frac{N_1}{N}d, \sigma^2\right)$ in the first and second groups, respectively, where $d = \frac{\mu_1 - \mu_2}{\sigma}$ denotes the effect size and σ^2 represents the common group variance. The identifiability constraint dictates that the global mean of the latent trait will be 0 for all N patients. The main objective of the RCT is the comparison of $H_0{:}d = 0$ and $H_1{:}d \neq 0$. Clinically, the two groups are said to differ if d is larger than a given effect size. For the observed variable, the classical sample size formula for a two-sided test, the size at α and the desired power at $1 - \beta$, is as follows:

$$N_1 = \frac{(r+1)\left(z_{\alpha/2+}z_\beta\right)^2}{rd^2}, \tag{2}$$

where r is the ratio of the sample size in group 2 to that in group 1 (i.e., group allocation ratio), $N_2 = rN_1$, and z_τ indicates the 100τth percentage of the cumulative normal distribution [2]. The latent trait is assumed to follow a Gaussian distribution; thus, this formula could also be suitable for sample size calculation based on the distribution of the latent trait. In practice, σ^2, μ_1, and μ_2 are unknown population parameters that characterize an unobserved latent variable; nevertheless, initial estimates from