ELSEVIER

Alzheimer's & Dementia: Translational Research & Clinical Interventions ■ (2017) 1-6

Alzheimer's & Dementia

Featured Article

Design of pilot studies to inform the construction of composite outcome measures

Steven D. Edland*, M. Colin Ard, Weiwei Li, Lingjing Jiang

■■■

Abstract

Background: Composite scales have recently been proposed as outcome measures for clinical trials. For example, the Prodromal Alzheimer's Cognitive Composite (PACC) is the sum of *z*-score normed component measures assessing episodic memory, timed executive function, and global cognition. Alternative methods of calculating composite total scores using the weighted sum of the component measures that maximize signal-to-noise ratio of the resulting composite score have been proposed. Optimal weights can be estimated from pilot data, but it is an open question as how large a pilot trial is required to calculate reliably optimal weights.

Methods: We describe the calculation of optimal weights and use large-scale computer simulations to investigate the question as how large a pilot study sample is required to inform the calculation of optimal weights. The simulations are informed by the pattern of decline observed in cognitively normal subjects enrolled in the Alzheimer's Disease Cooperative Study Prevention Instrument cohort study, restricting to $n = 75$ subjects aged 75 years and older with an ApoE E4 risk allele and therefore likely to have an underlying Alzheimer neurodegenerative process.

Results: In the context of secondary prevention trials in Alzheimer's disease and using the components of the PACC, we found that pilot studies as small as 100 are sufficient to meaningfully inform weighting parameters. Regardless of the pilot study sample size used to inform weights, the optimally weighted PACC consistently outperformed the standard PACC in terms of statistical power to detect treatment effects in a clinical trial. Pilot studies of size 300 produced weights that achieved near-optimal statistical power and reduced required sample size relative to the standard PACC by more than half.

Discussion: These simulations suggest that modestly sized pilot studies, comparable to that of a phase 2 clinical trial, are sufficient to inform the construction of composite outcome measures. Although these findings apply only to the PACC in the context of prodromal Alzheimer's disease, the observation that weights only have to approximate the optimal weights to achieve near-optimal performance should generalize. Performing a pilot study or phase 2 trial to inform the weighting of proposed composite outcome measures is highly cost-effective. The net effect of more efficient outcome measures is that smaller trials will be required to test novel treatments. Alternatively, second generation trials can use prior clinical trial data to inform weighting, so that greater efficiency can be achieved as we move forward.

Keywords: Alzheimer's disease; Phase 2 clinical trial; Phase 3 clinical trial; Composite endpoint; Cognitive decline; Secondary prevention; Power; Sample size

*Corresponding author. Tel.: ■■■; Fax: ■■■.
E-mail address: sedland@ucsd.edu

http://dx.doi.org/10.1016/j.trci.2016.12.004

1. Introduction

Composite endpoints have received increasing attention as potential outcome measures for clinical trials in Alzheimer's disease (AD). Composites can be defined as the sum of items taken from component instruments of a cognitive battery [1]. Or, more simply, composites can be defined as the sum of established cognitive instruments. One such composite is the Preclinical Alzheimer's Cognitive Composite or Prodromal Alzheimer's Cognitive Composite (PACC) [2]. The PACC is constructed from component measures assessing episodic memory, timed executive function, and global cognition and is the primary outcome measure for a major ongoing trial [3]. We have described how the performance of a composite endpoint depends on the weighting used and how optimal weights can be derived if the multivariate distribution of change scores on component measures is known [4]. The multivariate distribution of change scores of the component measures is typically not known but can be estimated if pilot data are available, for example, from a prior trial or from a prior representative registry study using the component instruments. An important consideration is whether prior data are sufficient to inform weighting parameters for a composite outcome measure and, in particular, how large sample size would be required to meaningfully inform calculation of weights. In this article, we use data from a completed registry trial to describe calculation of optimal weights and to investigate the question of what size pilot study is sufficient to inform calculation of optimal weights.

2. Methods

In overview, we use simulations informed by data from a completed registry trial, the Alzheimer's Disease Cooperative Study Prevention Instrument (PI) trial, to demonstrate optimal weighting and investigate the question as how large a pilot study is required to determine weights that improve the performance of the PACC. In the text that follows we briefly describe the PACC and the PI trial and then formally characterize optimal weights and computer simulation procedures.

2.1. Preclinical Alzheimer's Cognitive Composite

We use the PACC [2] to demonstrate the influence of weighting on characteristics of the composite scale. The PACC is a weighted sum of well recognized and validated component instruments, the Mini-Mental Status Examination (MMSE) assessing global cognition function [5], the Free and Cued Selective Reminding task (FCSRT) assessing episodic memory [6], and the WAIS-R Digit Symbol task (Digit Symbol), a timed test of processing speed and memory function [7], and the WMS-R Logical Memory story delayed recall task (Logical Memory) [8].

2.2. Prodromal AD PI cohort

Pilot study longitudinal data for the PACC to inform instrument behavior and clinical trial design are not available [2]. However, roughly comparable component instruments are available from the PI protocol conducted by the Alzheimer's Disease Cooperative Study [9]. The PI protocol performed annual neuropsychometric and functional assessments of 644 cognitively normal older persons (age 75 years and older). Although there was no randomization to treatment, the PI enrollment and assessment procedures mimicked that of a clinical trial, with primary purpose to assess the utility of the components of the assessment battery as potential endpoints for an Alzheimer prevention trial, and these data were used in the initial description of the PACC [2]. The PACC components that were not assessed in the PI study were the MMSE and the Logical Memory test. Comparable domain-specific instruments used in their stead were the modified MMSE [10] substituting for the MMSE, and the New York University Paragraph delayed recall test [11] substituting for the Logical Memory test. When the distinction is relevant, we call the resulting composite the PI-PACC to distinguish it from the PACC constructed from the MMSE, FCSRT, Digit Symbol, and Logical Memory test.

Donohue et al. [2] restricted their analysis to subjects with an ApoE E4 risk allele, and we follow suit. Subjects aged 75 years and older with this genetic risk profile have with high likelihood an underlying Alzheimer neurodegenerative process, and hence these subjects are an approximate representation of clinically normal, AD biomarker positive subjects that are the target of contemporary secondary prevention trials [2]. We call this subset of the PI cohort the PI Prodromal AD cohort. Baseline through month 36 data are available for 75 of these subjects (mean age at baseline 78.5 years [standard deviation 2.9 years], 59% female), and these longitudinal data are used to inform the simulations reported here. Q6

2.3. Optimal weights

We assume the primary analysis is mixed model repeated measure (MMRM) comparing change first to last in treatment versus change first to last in control [2]. To simplify presentation, we assume complete data for all simulations. Including missing values in simulations would reduce power given a total sample size, but would not appreciably impact the relative efficiency of trial designs and endpoints, which is the focus of this article. We further make the usual assumption that an effective treatment would shift the mean change but not affect the variability of change (constant variance of change in treatment and control arms). Under these assumptions, optimal weights for constructing a composite endpoint are a simple function of two sets of parameters, the expected change and the covariance of change of the component measures [4]. Given the vector μ of expected change scores of