ELSEVIER

Alzheimer's & Dementia: Translational Research & Clinical Interventions ■ (2016) 1-7

Alzheimer's & Dementia

Featured Article

Cross-validation of optimized composites for preclinical Alzheimer

Michael C. Donohue[a,*], Chung-Kai Sun[a], Rema Raman[a], Philip S. Insel[b,1], Paul S. Aisen[a], and NA-ADNI[1], AIBL[2], J-ADNI[3]

[a]Alzheimer's Therapeutic Research Institute, University of Southern California, San Diego, CA, USA
[b]University of Lund, Sweden; [c]University of California, San Francisco, CA, USA

Abstract

Introduction: We discuss optimization and validation of composite end points for presymptomatic Alzheimer clinical trials. Optimized composites offer hope of substantial gains in statistical power or reduction in sample size. But there is tradeoff between optimization and face validity such that optimization should only be considered if there is a convincing rationale. As with statistically derived regions of interest in neuroimaging, validation on independent data sets is essential.

Methods: Using four data sets, we consider the optimized weighting of four components of a cognitive composite which includes measures of (1) global cognition, (2) semantic memory, (3) episodic memory, and (4) executive function. Weights are optimized to either discriminate amyloid positivity or maximize power to detect a treatment effect in an amyloid-positive population. We apply repeated 5 × 3-fold cross-validation to quantify the out-of-sample performance of optimized composite end points.

Results: We found the optimized weights varied greatly across the folds of the cross-validation with either optimization method. Both optimization methods tend to down-weight the measures of global cognition and executive function. However, when these optimized composites were applied to the validation sets, they did not provide consistent improvements in power. In fact, overall, the optimized composites performed worse than those without optimization.

Discussion: We find that component weight optimization does not yield valid improvements in sensitivity of this composite to detect treatment effects.

Keywords: Preclinical Alzheimer; Cognitive composites; End-point validation

1. Introduction

Cognitive composites are weighted sums of component cognitive assessments. For example, the preclinical Alzheimer cognitive composite (PACC) [1] is a weighted sum of four components: (1) Free and Cued Selective Reminding Test (FCSRT); (2) Logical Memory Paragraph Recall; (3) Mini–Mental State Examination (MMSE); and

[1]Data used in preparation of this article were obtained from the (North American) Alzheimer's Disease Neuroimaging Initiative (NA-ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of NA-ADNI investigators can be found at http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.[2] Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle flagship study of ageing (AIBL) funded by the Commonwealth Scientific and Industrial Research Organisation (CSIRO) which was made available at the ADNI database (www.loni.usc.edu/ADNI). The AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at www.aibl.csiro.au.[3]Data used in this research was originally obtained by Japanese Alzheimer's Disease Neuroimaging Initiative http://humandbs.biosciencedbc.jp/en/hum0043-v1 (led by Prof. Takeshi Iwatsubo) and available at the website of the National Bioscience Database Center (NBDC; http://biosciencedbc.jp/en/) of the Japan Science and Technology Agency (JST).

*Corresponding author. Tel.: ■■■; Fax: ■■■.
E-mail address: mdonohue@usc.edu

http://dx.doi.org/10.1016/j.trci.2016.12.001

(4) Digit Symbol Substitution Test. The components were chosen, based on a broad literature review, for their sensitivity to decline in preclinical and prodromal stages of Alzheimer disease. For example, the MMSE has demonstrated sensitivity to decline preclinical Alzheimer populations [2–4]. In its current implementation, PACC components are weighted equally, with the aim of giving more than half of the total weight to episodic memory (components 1, 2, and part of 3, but also giving importance to orientation and language (parts of component 3) and executive function (component 4).

The PACC has been criticized on several fronts. It has been suggested that MMSE has a restricted range of likely scores in this population and should be dropped from composite measures for preclinical Alzheimer [5]. Others have suggested a more data-driven approach should be used to select components and weights should be optimized to increase power to detect treatment effects or reduce required sample size [6]. Our motivation is to explore the out-of-sample performance of versions of the PACC with such optimized component weights.

The component weights can be optimized according to any reasonable criterion, for example, to maximize placebo group decline [6], or maximize power, or to minimize the smallest detectable effect size. All optimization algorithms are "greedy" in the sense that their solution is guaranteed to be optimal only for the given training set, and this tends to come at the cost of generalizability to new data. Cross-validation [7] can be used to provide an assessment of out-of-sample performance.

2. Methods

2.1. Data sets

We explore composite optimization in cohorts with normal cognition from four studies: (1) North American Alzheimer's Disease Neuroimaging Initiative (NA-ADNI [8]), (2) Japan-ADNI (J-ADNI [9]), (3) Australian Imaging, Biomarkers and Lifestyle Flagship Study of Ageing (AIBL [10]), and (4) Alzheimer's Disease Cooperative Study Prevention Instrument (ADCS-PI [1]). For each data set, we consider a "target" population (e.g., Aβ+, APOE ε4+, or clinical dementia rating global [CDR-G] progressors) and a complementary "reference" population (e.g., Aβ−, APOE ε4−, or CDR-G stable). Table 1 summarizes the composite components available in the four data sets and the target/reference groups used. For this analysis, we use the total free recall score from the FCSRT in the ADCS-PI study.

2.2. Composite construction

The PACC is the sum of the four component z-scores, defined

Table 1
External validation of weights optimized using AIBL

	AIBL ($\widehat{\mathbf{w}}$)	NA-ADNI	J-ADNI	ADCS-PI	
Grouped by	PET	PET/CSF		APOE ε4	CDR-G
z_1 MMSE	MMSE (6%)	MMSE		3MSE	
z_2 FCSRT	CVLT (55%)	ADAS-COG		FCSRT	
z_3 LM	LM (35%)	LM		NYU	
z_4 Digit	Digit (5%)	Digit		Digit	
δ (equal $\widehat{\mathbf{w}}$)	33%	42% (year 2)	35%	48%	14%
δ (logistic $\widehat{\mathbf{w}}$)	27%	*	54%	95%	15%

Abbreviations: AIBL, Australian Imaging, Biomarkers and Lifestyle; ADNI, Alzheimer's Disease Neuroimaging Initiative; NA-ADNI, North American ADNI; J-ADNI, Japan-ADNI; ADCS-PI, Alzheimer's Disease Cooperative Study Prevention Instrument; CDR-G, clinical dementia rating global; MMSE, Mini–Mental State Examination; 3MSE, modified MMSE; FCSRT, Free and Cued Selective Reminding Test; CVLT, California Verbal Learning Test; ADAS-Cog, Alzheimer's Disease Assessment Scale–Cognitive; LM, Logical Memory; NYU, New York University Paragraph Recall; Digit, digit symbol substitution; PACC, preclinical Alzheimer cognitive composite.

NOTE. The MMSE, FCSRT, LM, and digit rows represent the four components of the PACC. Columns 2 through 6 represent the four pilot data sets, and indicated groupings, used to explore the performance of the PACC. The indicated proxy components (e.g., CVLT) were used when the actual PACC components (e.g., FCSRT) were not available in a study (e.g., AIBL). To explore optimized weighting of the PACC, we fit AIBL data to a logistic model of Aβ+ status with month 36 component change z-scores as covariates. The regression coefficients from this model (rescaled to sum to 100%) provide a weighting tuned to discriminate Aβ+ status. The resulting weights are in bold and parentheses in the AIBL column, and the resulting minimum detectable δ is summarized in the bottom row. The numerically minimized δ was 25% (2% smaller than the logistic-derived δ), but this required weighting digit in the opposite direction (6% MMSE, 48% CVLT, 54% LM, and −8% digit).

*The AIBL-optimized PACC was not significantly different at any visit in ADNI, whereas the original was significant only at year 2.

$$z_{jt} = \frac{(y_{jt} - y_{j0})}{\sigma_{j0}},$$

for component $j = 1,\ldots,4$ at time t, where σ_{j0} is standard deviation of component score y_{j0}. We consider optimized versions of the PACC which are weighted sums:

$$Y_t(\mathbf{w}) = z_{1t}w_1 + z_{2t}w_2 + z_{3t}w_3 + z_{4t}w_4,$$

where w = (w_1, w_2, w_3, w_4) is the weight vector or list of the four component weights. We orient each composite the same way (e.g., lower scores denote worsening) and constrain the weights to sum to one. The originally proposed PACC uses equal weights, effectively: $w_1 = w_2 = w_3 = w_4 = 0.25$.

2.3. Optimization

The feasibility of using the PACC to detect treatment effects in an elderly population with preclinical Alzheimer (normal cognition but abnormal amyloid accumulation in brain) was based on natural history data such as that depicted in Fig. 1. Change is estimated in the amyloid-β (Aβ) positive and negative groups, and the smallest detectable treatment