Keck School of Medicine of USC

Alzheimer's Therapeutic Research Institute

Cross-Validation of Optimized Composites for Preclinical Alzheimer's

Michael C. Donohue[1], Chung-Kai Sun[1], Rema Raman[1], Philip Insel[2,3], Niklas Mattson[2], Beth Mormino[4], Reisa Sperling[4], Paul S. Aisen[1] for ADNI, AIBL, and J-ADNI

[1] Alzheimer's Therapeutic Research Institute, University of Southern California; [2]University of Lund; [3]University of California, San Francisco; [4]Havard Medical School

BACKGROUND

Cognitive composites are weighted sums of component assessments. For example, the **Preclinical Alzheimer's Cognitive Composite (PACC)** (Donohue et al 2014) is a weighted sum of four components:

1. Free and Cued Selective Reminding Test (FCSRT)
2. Logical Memory Paragraph Recall
3. Mini Mental State Exam (MMSE)
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 the disease. 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 component weights can be "**optimized**" according to any reasonable criterion, e.g. to maximize power or to minimize the smallest detectable effect size. All optimization algorithms are "*greedy*" – their solution is optimal for the given *training set* but may not perform well on other independent datasets. Cross-validation can be used to assess this out-of-sample performance.

METHODS

Datasets. We explore composite optimization in cohorts with normal cognition from four studies: (1) North American Alzheimer's Disease Neuroimaging Initiative (NA-ADNI), (2) Japanese-ADNI (J-ADNI), (3) Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL), and (4) Alzheimer's Disease Cooperative Study Prevention Instrument (ADCS-PI). For each dataset we consider a "**target**" population (e.g. Aβ+, APOEε4+, or Clinical Dementia Rating Global [CDR-G] progressors) and a complementary "**reference**" population (e.g. Aβ-, APOEε4-, or CDR-G stable) (Figure 1). Table 1 summarizes the composite components available in the four datasets and the target/reference groups used.

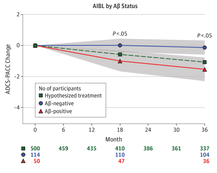

Figure 1. The PACC in AIBL. Aβ group profiles and the smallest detectable effect, $\boldsymbol{\delta}$ (green), based on AIBL data. The sample size calculation is based on MMRM and assumes 80% power, 5% two-sided, 3 year trial, n=500 per group, and 30% attrition (Donohue et al 2014).

Composite Construction. The PACC is the sum of the four component z-scores, defined:

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

for component j (j=1,…,4) at time t, where σ_{j0} is the standard deviation of the j^{th} component at baseline. We considered 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 $\mathbf{w}=(w_1,w_2,w_3,w_4)$ is the weight vector ($w_j \geq 0$ and $w_1 + w_2 + w_3 + w_4 = 1$).

Optimization. We derive optimized weights using two approaches:

1. **Logistic regression:** Weights are derived from a logistic regression to discriminate the target (e.g. Aβ+) from the reference population (e.g. Aβ-) based on three-year component change scores, and
2. **Minimize δ:** Weights are derived to minimize the detectable treatment effect (δ) as a percentage of the group difference in change from baseline between the target and reference populations. These weight are found by submitting the sample size formula to a numerical optimization routine (Nelder & Mead 1965).

External validation. To explore optimized weighting of the PACC, we first attempted a simple external validation of the weights derived from AIBL. Power calculations are conducted using the AIBL-optimized composite applied to the other datasets, and the resulting minimum detectable effect size is reported for each data set and optimization method.

Cross-validation. Figure 2 provides a basic schematic of cross-validation. With each fold of the cross-validation, we derive the optimal weights using the training set, then test its performance (minimum detectable δ) on the validation set. Due to small sample sizes, we used a relatively small number of folds (3). We used repeated cross-validation (Burman 1989), which is akin to applying the bootstrap to cross-validation. We repeat each 3-fold cross-validation 5 times ("5x3 cross-validation"), yielding a total of 15 estimates of $\mathbf{w}$ and associated δ per data set. We graphically summarize the median and range of these estimates by data set, as well as the pooled medians.

Power calculations assume a Mixed Model of Repeated Measures to estimate treatment effect at 36 months, 6-month visit intervals, 500 participants per group, 30% attrition, 5% α, and 80% power.

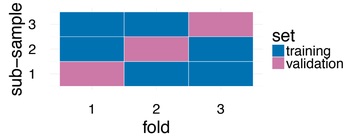

Figure 2. With k-fold cross-validation we split the sample up into k non-overlapping sub-samples of size n/k which take turns as the validation set (red), leaving the remaining sample (blue) for training set. In our case, weights are optimized on each training set and out-of-sample performance (minimum detectable δ) is assessed on validation sets.

RESULTS

Study (grouping)	A4 (Amyloid PET)	AIBL (Amyloid PET)	NA-ADNI (PET/CSF)	J-ADNI (Amyloid PET)	ADCS-PI (APOEε4)	ADCS-PI (CDR-G)
Component z_1	MMSE	MMSE (**6%**)	MMSE	MMSE	3MSE	3MSE
Component z_2	FCSRT	CVLT (**55%**)	ADAS-Cog	ADAS-Cog	FCSRT	FCSRT
Component z_3	LM	LM (**35%**)	LM	LM	NYU	NYU
Component z_3	Digit	Digit (**5%**)	Digit	Digit	Digit	Digit
δ using equal weights		33%	42% (year 2)	35%	48%	14%
δ using logistic regression weights		27%*	†	54%	95%	15%

Table 1. External validation of weights optimized using 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 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 minimum possible (numerically optimized) δ was 25%, but this required weighting Digit Symbol in the wrong direction.
† The logistic optimized PACC was not significantly different at any visit in ADNI, while the original was significant only at year 2.

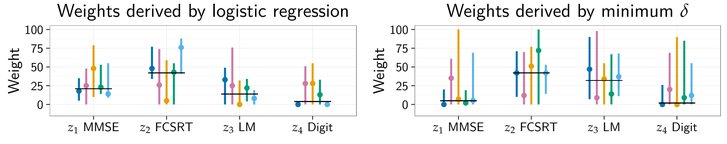

Figure 3. Optimized composite weights across the cross-validation subsamples. Medians (dots) and range (vertical lines) of the logistic-optimized (left) and minimum detectable $\boldsymbol{\delta}$ (right) by data set (colors) across the 15 repeated cross-validation subsamples (see color legend in Figure 4).

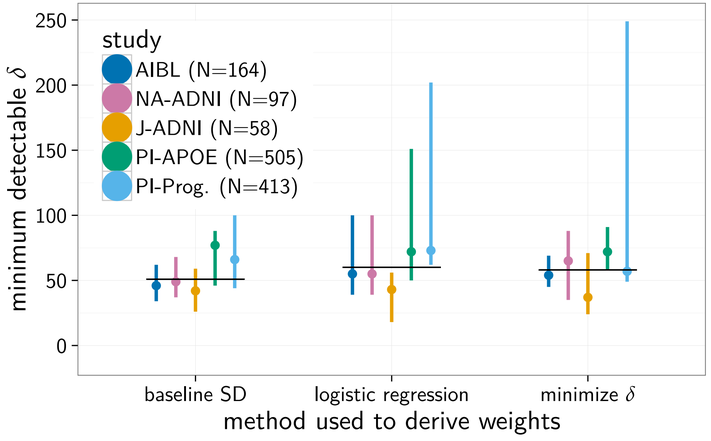

Figure 4. Minimum detectable δ attained out-of-sample. Medians (dots) and range (vertical lines) of the minimum detectable δ attained out-of-sample using each of the 3 indicated weighting methods.

DISCUSSION

Summary

- There is tradeoff between optimization and face validity such that optimization should only be considered if there is a convincing rationale.
- Component weight optimization does not yield valid improvements in sensitivity of PACC to detect treatment effects. We did not explore item level optimization, which would be even more costly to face validity.
- Both MMSE and Digit Symbol Substitution were consistently down-weighted by optimization. However, down-weighting them did not reliably improve composite performance out-of-sample. MMSE has good face validity as a global assessment and has demonstrated sensitivity to preclinical decline (Amieva, et al 2008). Digit Symbol Substitution has good face validity as a measure of executive function. We do not find adequate justification to omit either the MMSE nor Digit Symbol Substitution from the PACC.

Limitations of external validation:

- It is expensive to conduct additional validation studies.
- Existing data is never ideally matched with respect to the study populations or assessments.

Limitations of cross-validation:

- Mere simulation of real-world out-of-sample replication.
- Sub-samples may not be of sufficient size to generate reliable weight estimates.

REFERENCES

1. Donohue, MC, et al. (2014). The preclinical Alzheimer cognitive composite: measuring amyloid-related decline. *JAMA Neurology*, *71*(8), 961-970.
2. Burman, P. (1989). A comparative study of ordinary cross-validation, v-fold cross-validation and the repeated learning-testing methods. *Biometrika*,*76*(3), 503-514.
3. Nelder, JA, Mead R. (1965) A simplex method for function minimization. *The Computer Journal*, 7(4), 308-313.
4. Amieva, H, et al. (2008) Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. *Annals of Neurology* 64(5), 492-498.

ACKNOWLEDGEMENTS

We are grateful to the study teams and study participants of the ADCS-PI, NA-ADNI, J-ADNI, and AIBL studies.