Using imputation to harmonize longitudinal measures of cognition across two large cohorts: AIBL and ADNI

Rosita Shishegar[1], Timothy Cox[1], David Rolls[1], Vincent Dore[1], Fiona Lamb[2], Joanne Robertson[3], Simon Laws[4], Tenielle Porter[4], Paul Maruff[5], Greg Savage[6], Christopher Rowe[2], Colin Masters[3], Mike Weiner[7], Victor Villemagne[3], Samantha Burnham[1] for the Alzheimer's Disease Neuroimaging Initiative8 and the AIBL Research[9]

1. The Australian e-Health Research Centre, CSIRO, Melbourne, Australia; 2.Department of Medicine, Austin Health, Heidelberg, VIC, Australia; 3.Florey Institute of Neuroscience and Mental Health, The University of Melbourne, Parkville, VIC, Australia; 4.School of Biomedical Sciences, Faculty of Health Sciences, Curtin University, Bentley, WA, Australia; 5.Cogstate Ltd, Melbourne, VIC, Australia; 6.ARC Centre of Excellence in Cognition and its Disorders, and Department of Psychology, Macquarie University, Sydney, NSW, Australia; 7.Center for Imaging of Neurodegenerative Diseases, University of California-San Francisco, San Francisco, CA, USA.

HEALTH & BIOSECURITY
www.csiro.au

Backgrounds

To ensure the generalisability of findings and consider more nuanced hypotheses, larger sample sizes are required. Combining data from different but similar study cohorts is one solution. However, the disparity of these datasets, e.g. using differing tests to assess specific cognitive domains, makes this a non-trivial task. Here, we propose a harmonisation solution using **imputation strategies**[1] for **cognitive memory performance in AIBL[2] and ADNI[3]**.

Methods

Data harmonization steps included: 1) data cleaning 2) single naming convention 3) harmonizing data types 4) joining datasets 5) imputing unmeasured test scores (e.g. CVLT-II for AIBL and RALVT for ADNI participants)

Data imputation model set up:

The joined data was assembled in the long format so that each row is a single time point per subject.

The final joined data set included 36 columns. The columns included: a) demographic measurements: age, gender, years of education, clinical classification (NC, MCI and AD), a genetic risk factor (carriage of the apolipoprotein E ε4 (APOE-ε4) allele, and indicator of the source dataset (AIBL, ADNI); b) clinical tests: MMSE and CDR; c) 14 cognitive test scores and their time dependent variables.

Non-parametric multivariate imputation using random forests (missForest)[4] was employed to impute data.

Results

Table 1: Demographics table for AIBL, ADNI and the joint dataset.

	AIBL dataset (N=1805)	ADNI dataset (N=2122)	Joined dataset (N=3927)	Statistic (df)	P-value
Clinical Classification NC/MCI/AD [N]	1180/ 297/ 328	791/ 962/ 369	1971/ 1259/ 697	$\chi^2(2)$= 407.5	<0.001
Gender: Males [N (%)]	777 (43.0)	1129 (53.2)	1906 (48.5)	$\chi^2(1)$= =39.881	<0.001
Age (years) [mean (sd)]	72.5 (7.72)	73.3 (7.21)	73.0 (7.46)	t(3728.5)= -3.3021	<0.001
APOE	941/ 456/ 91	1114/ 739/ 194	2055/ 1195/ 285	$\chi^2(2)$=31.192	<0.001
Mounts of Follow-up [Mean (sd)]	45.2 (34.8)	37.5 (35.7)	41.1 (35.5)	t(3853.3)= 6.7951	<0.001
Years of Education [N] <9 / 9-12 / 13-15 / >15	189 / 663 / 345 / 551	23 / 288 / 401 / 1410	212 / 951 / 746 / 1961	χ2(3)= 628.06	<0.001

	MAE	
	LMII	MMSE
ADNI data imputed using ADNI	1.86	1.42
ADNI data imputed using AIBL	2.34	2.00

Table 2: The mean absolute error (MAE) of imputed LMII and MMSE scores compared to actual values for a subset of ADNI dataset used as the test data. The rest of the ADNI data as well as the AIBL data were used to train the imputation models.

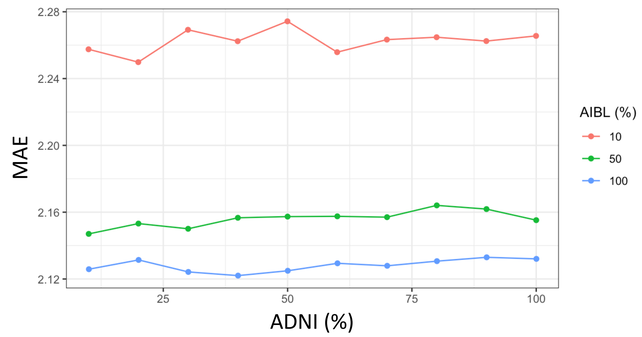

Figure 1: Performance of imputed simulated missing ADNI LMII scores with **different size of training and missing data**. The performance is calculated using the **mean absolute error** (MAE) of imputed and actual data. Different sizes of training data of 10%, 50%, and 100% of AIBL dataset and different sizes of simulated missing data samples of ADNI data set (equal to the size of 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, and 100% of AIBL dataset) were used.

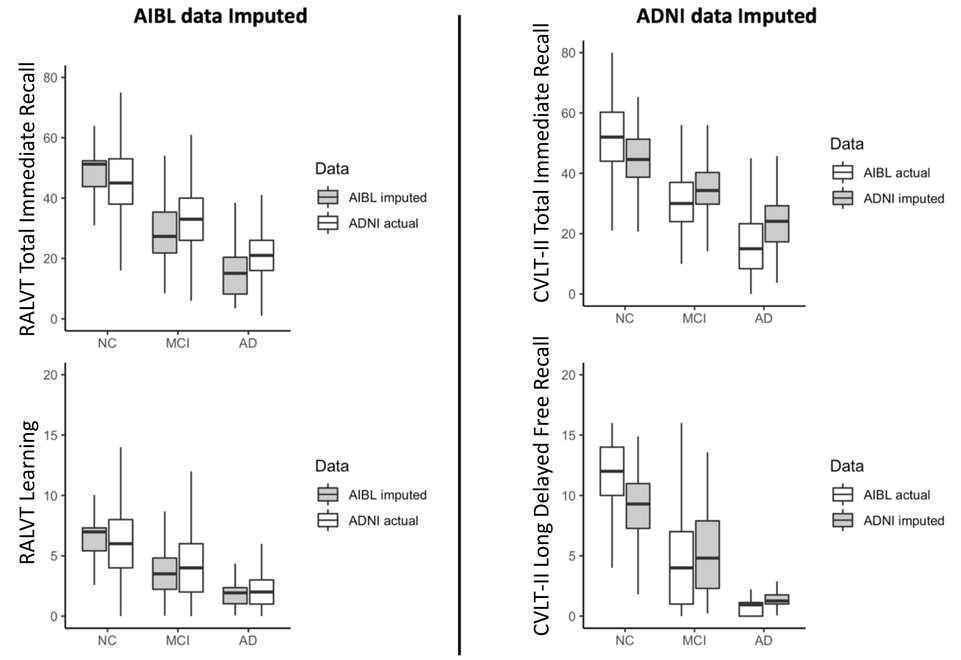

Figure 2:
AIBL data imputed: Distribution of the ADNI RALVT scores compared with the imputed AIBL RALVT scores.
ADNI data imputed: Distribution of the AIBL CVLT-II scores compared with the imputed ADNI CVLT-II scores. Results show **high levels of significant discrimination** (p<.001) between clinical classifications for the actual and the imputed scores of the CVLT-II and RAVLT.

Discussion and Conclusions

Although there are significant differences between demographic scores of two datasets (Table 1), the proposed imputation method predicts the unmeasured test scores with a high accuracy that is comparable to the accuracy of missing data imputation in a single dataset (Table 2).

The results here suggest it is possible to use data imputation, capitalising on underlying structures and relationships, to impute specific tests scores in a cohort for which that test what not administered. In turn, providing a practical solution for data harmonization across large, longitudinal datasets.

FOR FURTHER INFORMATION

Rosita Shishegar
e rosita.shishegar@csiro.au
w www.csiro.us

ACKNOWLEDGEMENTS

AIBL is a large collaborative study and a complete list of contributors can be found at our website www.aibl.csiro.au

We thank all who took part in the study.

References

1doi:10.1038/srep21689. 2doi:10.1017/S1041610209009405. 3doi:10.1212/WNL.0b013e3181cb3e25. 4doi:10.1093/bioinformatics/btr597.