Predicting Child Language Outcomes Across Diverse Longitudinal Cohorts: A Machine Learning Approach

Marja Laasonen*, Rosa González Hautamäki, Federico Màlato, Jade Plym, Sini Smolander, Eva Arkkila, Pekka Lahti-Nuuttila, Sari Kunnari, Penny Levickis, Cristina McKean, & Patricia Eadie
* *University of Eastern Finland, Finland*
marja.laasonen@uef.fi

Paper Abstract

Introduction: Developmental language disorder (DLD) is a highly prevalent condition that significantly impacts children's language development, academic success, and overall well-being. Despite extensive research efforts, we do not fully understand the factors influencing typical language development and the emergence of DLD. This study employs machine learning techniques to predict and examine the generalisability of child language outcomes across two diverse longitudinal cohorts: the Helsinki Longitudinal Specific Language Impairment Study (HelSLI) in Finland and the Early Language in Victoria Study (ELVS) in Australia.

Specifically, we asked: (1) Can group membership (typical development, TD vs. low language outcome, LLO) be predicted using a comprehensive set of cognitive/neuropsychological variables, including linguistic measures, in the ELVS dataset? (2) Which cognitive/neuropsychological and linguistic variables are the most informative predictors of group membership in the ELVS dataset? (3) How do the predictive models and critical variables identified in the HelSLI study compare to those from the ELVS dataset, and what does this reveal about the generalisability of findings across populations?

Methods: We employed an ensemble machine learning approach for two-class classification problems, specifically Random Forest (RF). RF provides an accurate model that explores insights into which variables contribute most toward predictive accuracy.

Results: The previous cross-sectional HelSLI studies showed that using a machine learning approach, neuropsychological and linguistic variables could accurately (85–90 %) classify TD and DLD in 3–7-year-old monolingual and sequentially bilingual children. Among the neuropsychological set, language and verbal memory were most important in classifying monolingual and bilingual children. In the linguistic set, the best classifiers differed between the language groups, with variables corresponding to language production being superior in classifying monolingual groups and language comprehension in bilingual groups. The results with the ELVS dataset indicate how similar variables contribute to the prediction of group membership.

Discussion: The findings of this study have important implications for our understanding of the complex, multifactorial nature of DLD and the development of more effective, culturally sensitive approaches to early identification and intervention. By comparing the predictive models and key variables across the HelSLI and ELVS datasets, we aim to

identify universal and language-specific factors that shape language development and contribute to the emergence of language disorders.

Keywords: cognitive/neuropsychological factors, cross-linguistic, developmental language disorder, low language outcome, machine learning