

A Robust Feature Selection Analysis of Early Childhood Development Predictors in Thailand Using Stability Selection

Saengrawee Sutas¹, Wasurut Ployluan², Thanaporn Patikorn³

¹Lecturer, Boromarajonani College of Nursing, Chonburi, Faculty of Nursing, Praboromarajchanok Institute, Thailand

²Public Health Technical Officer, Professional Level, Regional Health Promotion Center 6, Chonburi, Thailand

³Lecturer, Faculty of Science and Technology, Rajamangala University of Technology Suvarnabhumi, Thailand

How to Cite: Sutas, S., Ployluan, W., & Patikorn, T. (2026). A Robust Feature Selection Analysis of Early Childhood Development Predictors in Thailand Using Stability Selection. *International Journal of Child Development and Mental Health*, 14(1), 53-66.

*Corresponding author:

Email: saengrawee@bnc.ac.th

Article Info:

Received: 24 October 2025

1st Revision: 4 December 2025

Accepted: 13 January 2026

Abstract

Early identification of developmental delays is critical for timely intervention. However, the specialized training and time required for frequent professional evaluations create a high clinical demand. This highlights the need for proxy predictors, features that are strong predictors of developmental delays that are also easily observable by caregivers, which can act as early triggers for professional referral. This study aimed to identify robust clinical and environmental predictors of suspected developmental delay across five DSPM domains and overall development among children in Chonburi, Thailand. Using a cross-sectional dataset (N=300) with 243 engineered features, a two-stage statistical pipeline was employed: (1) Stability Selection (100 iterations, 50% subsampling) to identify variables with a selection frequency ≥ 0.50 ; and (2) Non-parametric Bootstrapping (1,000 iterations) with multivariate logistic regression to calculate Median Odds Ratios (OR) and 95% Confidence Intervals (CI). Dental plaque emerged as a primary predictor, significantly increasing the risk for overall developmental delay (OR = 5.36) and linguistic delays. Other stable risk factors included starting electronic media between ages 2.0–2.9 years (OR = 6.95), a history of childhood pneumonia (OR = 5.96), and receiving the DSPM manual only during vaccination visits (OR = 5.47). Conversely, avoiding electronics (OR = 0.29) and proactive DSPM utilization (OR = 0.30) were significant protective factors for expressive language. No stable predictors reached the threshold for motor or social domains. Findings emphasize the sentinel role of caretakers' health literacy, oral health, and digital hygiene in developmental screening. Poor oral health, early digital device usage, and lack of health literacy through DSPM are proxy predictors for developmental delay, while caretakers' health literacy through utilization of DSPM and avoiding usage of electronic devices lowers the likelihood of developmental delay. Pediatric nursing interventions should prioritize caretakers' health literacy and shift from reactive to proactive DSPM utilization to improve developmental outcomes.

Keywords:

Early Childhood Development,
Developmental Screening,
Caregiver-Observable Predictors,
Feature Selection Analysis, DSPM.

Introduction

Early childhood development is a critical foundation that influences a person's long-term health and success. Research indicates that physical growth and cognitive progress during the first two years of life are linked to intelligence and academic achievement in later years (Bornstein et al., 2013A; Bornstein et al., 2013B; Räikkönen et al., 2009). These effects are examples of “developmental cascades” where events, such as traumas, or development in one area, such as early fine motor skills, can impact development in other areas and/or development later in their lives. These effects suggest that early maturity in areas like motor skills provides a base for complex learning. Therefore, identifying developmental delays early is essential so that caregivers and health workers can intervene and ensure every child has a healthy developmental path. In Thailand, the Ministry of Public Health responded to rising rates of developmental delay by creating specialized screening tools (Jantanavivat et al., 2022). Since 2015, the primary tool used nationwide for children aged 0–5 is the Developmental Surveillance and Promotion Manual, or DSPM (Ministry of Public Health, 2020). The DSPM monitors growth at key milestones across five essential domains: Gross Motor (GM), Fine Motor and Adaptive (FM), Receptive Language (RL), Expressive Language (EL), and Personal and Social (PS) skills.

While the DSPM is the national “gold standard” for assessment, it is often impractical to evaluate every child regularly because the process is time-consuming and requires specific training. Consequently, there is a demand for “proxy predictors” which are strong predictors of developmental delays that are easier for parents or teachers to observe without professional training. Previous studies have highlighted lifestyle factors, such as excessive screen time and child temperament, as significant indicators of potential delay (Maneekunwong et al., 2025; Rithipukdee & Kusol, 2022). Identifying a broader range of robust predictors would allow for more frequent monitoring and earlier professional referrals. While current literature has identified factors such as screen time and child temperament, there remains a research gap regarding a comprehensive set of robust, stable predictors that can be reliably used across different milestone ages. To address this, evidence

is needed from regions that reflect contemporary Thai developmental trends. Chonburi Province, a primary economic and urban-industrial hub under the jurisdiction of Regional Health Center 6, provides a diverse and representative population for this inquiry.

Objective

This study aimed to identify the most reliable predictors of childhood development using a cross-sectional dataset of 300 participants in Chonburi, Thailand. Potential predictors were analyzed against six target variables: the five DSPM domains and one overall developmental score. To ensure the reliability of findings, stability selection was used to select the stable predictors. Logistic regression was then repeatedly fitted using stable predictors with bootstrapping to calculate Odds Ratios (OR) and confidence intervals (CI).

Literature Review

Early childhood is a foundational period that significantly influences long-term physical and cognitive outcomes. Research has established that physical growth and weight gain during the first two years of life are reliable predictors of intellectual abilities in young adulthood, with a decrease of 1-SD in test score per SD in growth (Räikkönen et al., 2009). This relationship is often described as a “developmental cascade,” where early maturity in specific areas, such as infant motor exploration at five months, predicts higher academic achievement during adolescence (Bornstein et al., 2013A). Another 14-year longitudinal study (Bornstein et al., 2013B) showed a similar effect of early cognitive development on later academic achievements. Longitudinal studies further suggest that early cognitive advantages often persist, with preschool-age IQ levels correlating strongly with educational success and cognitive stability into early adulthood both in Germany (Schneider et al., 2014) and in China (Zhu et al., 2019). In response to high rates of developmental delay, around 30% from 1999–2014, the Thai Ministry of Public Health established standardized tools for screening and promotion in 2015 called Developmental Surveillance and Promotion Manual, or DSPM (Jantanavivat et al.,

2022), which was updated in 2020 (Ministry of Public Health, 2020). DSPM has since become the primary instrument used nationwide for monitoring child development progress, especially at the milestone ages of 9, 18, 30, and 42 months across five domains (Thaineua et al., 2024). Recent evaluations indicate that while screening rates have increased, the prevalence of suspected developmental delay remains a concern (Thaineua et al., 2024). Studies in regional Thailand have specifically identified environmental and behavioral factors, such as excessive daily screen time and certain child temperaments, as significant risk factors associated with language and overall developmental delays (Rithipukdee & Kusol, 2022). A case-controlled study by Maneekunwong (Maneekunwong, 2025) also confirmed the effect of electrical device usage on suspected developmental delay in rural Thai children. Health-related datasets often contain a large number of variables, also called high dimensionality, spanning various aspects of each participant or patient. Identifying reliable predictors from high dimensional data is often challenging for traditional statistical methods, especially where the number of potential variables exceeds the sample size. Traditional methods, such as the Lasso penalty, can be sensitive to small changes in the data; however, Stability Selection addresses this by providing a more robust framework for variable selection (Meinshausen & Bühlmann, 2010). This approach has proven effective in various clinical settings, including the prediction of patient risk from electronic health records and the identification of biological markers (Zhou et al., 2013; Wehrens et al., 2011; Das et al., 2026). Furthermore, simulation studies suggest that Stability Selection offers superior control over false discovery rates compared to standard regularization methods, making it a more suitable choice for epidemiology and wide datasets where identifying truly impactful predictors is essential (Hyde et al., 2022).

Methodology

This cross-sectional study utilized a sample of 300 participants from Chonburi Province, conducted as part of a regional child development assessment by

Regional Health Center 6 (Chonburi), Department of Health, Ministry of Public Health. The sample was stratified into five age groups, namely 9, 18, 30, 42, and 60 months, with 60 children per cohort. For this study, 'age' refers to the full milestone month only (e.g., 9 months 0–29 days), with no participants recruited from intervening months. These specific intervals were selected to align with the “milestone ages” defined by the Developmental Surveillance and Promotion Manual (DSPM). This demographic provided a critical look at developmental outcomes within a region transitioning through rapid industrial and urban expansion, which may introduce specific environmental and socioeconomic determinants for pediatric health.

The dataset comprised 72 variables categorized into demographic, clinical, and developmental domains. Initial socio-demographic data included identification and location (*No*, *Province*, and *District*), followed by parental characteristics (*AgeMother*, *EduMother*, *OccMother*, *AgeFather*, *EduFather*, and *OccFather*) and primary caregiver profiles (*WhoCaregiver*, *AgeCaregiver*, *SexCaregiver*, *EduCaregiver*, and *OccCaregiver*). Household stability was assessed through *SufficientIncome*, *Marital Status*, and *FinancialSupport*. Maternal health and pregnancy history were captured via *AgeAtPregnancy*, *AgeAtANCI*, *Infection*, *InfectionMother*, *Supplements*, and maternal hematocrit levels (*HematocritMother1* and *HematocritMother2*). Child-specific health and nutritional data included *AgeAtScreening*, *BirthPlaceProvince*, *BirthPlaceHospital*, *SexChild*, and *ChildEmotions*, as well as clinical history variables (*DiseaseChild*, *DiseaseChildWhat*, *DiseaseChildWY1*, *DiseaseChildWY*, *SymptomChildWY1*, *HematocritChild1*, and *HematocritChild2*). Nutritional practices and supplementation were recorded through *Breastfeeding*, *BreastfeedingWhat*, *BreastfeedingNow*, *Folic*, and *FolicTake*. The home environment and behavioral stimulation were measured by engagement in *ReadingBook*, *Singing*, and *PlayingWithChild*, alongside sedentary behaviors such as *WatchingTV*, *WatchingTVWeekday*, *ElectronicStart*, *ElectEquip*, and *ElectEquipHour*. The study also evaluated the implementation of the Developmental Surveillance and Promotion Manual (DSPM) through variables including *DSPMReceived*,

WhenGetDSPM, StaffTeach, StaffTeachUnderstand, ValunteerTeach, ValTeachUnderstand, Read, Playing, Use, and Satisfaction. Physical health and growth outcomes were represented by *WtforAge, HtforAge, and WeightforHeight*, together with dental health indicators (*OralHealth, DentalCaries1, DentalCaries2, and DentalPlaque*).

Finally, the six primary target variables assessing developmental outcomes were *OverallDevelopment, GrossMotor, FineMotorAdaptive, Expressive Language, ReceptiveLanguage, and PersonalSocial*. All target variables were collected according to the DSPM. Descriptive statistics for each variable are presented in Table 1.

Table 1: Descriptive Statistics of Binary Variables

Variable	Category 1: N (%)	Category 2: N (%)	Unknown/ Missing: N (%)
SexCaregiver	Male: 14 (4.67%)	Female: 286 (95.33%)	
FinancialSupport	Not received: 68 (22.67%)	Received: 229 (76.33%)	3 (1%)
Infection	No/None: 290 (96.67%)	Have: 10 (3.33%)	
Supplements	Not received: 7 (2.33%)	Received: 293 (97.67%)	
BirthPlaceHospital	Private hospital: 24 (8%)	Government hospital: 276 (92%)	
SexChild	Male: 142 (47.33%)	Female: 158 (52.67%)	
DiseaseChild	No/None: 285 (95%)	Have: 15 (5%)	
DiseaseChildWY1	Ever/Used to: 82 (27.33%)	Never: 218 (72.67%)	
Breastfeeding	Eat: 266 (88.67%)	Do not eat: 34 (11.33%)	
BreastfeedingNow	Eat: 27 (9%)	Do not eat: 239 (79.67%)	34 (11.33%)
Folic	Not received: 150 (50%)	Received: 150 (50%)	
ReadingBook	No/None: 149 (49.67%)	Have: 150 (50%)	1 (0.33%)
Singing	No/None: 115 (38.33%)	Have: 185 (61.67%)	
PlayingWithChild	No/None: 18 (6%)	Have: 282 (94%)	
WatchingTV	Do not watch: 135 (45%)	Watch: 165 (55%)	
ElectEquip	Play: 140 (46.67%)	Do not play: 159 (53%)	1 (0.33%)
DSPMReceived	Not received: 85 (28.33%)	Received: 215 (71.67%)	
StaffTeach	Explained: 180 (60%)	Did not explain: 35 (11.67%)	85 (28.33%)
ValunteerTeach	Explained: 41 (13.67%)	Did not explain: 174 (58%)	85 (28.33%)
Read	Not yet read: 59 (19.67%)	Already read: 156 (52%)	85 (28.33%)
DentalPlaque	No/None: 171 (57%)	Have: 118 (39.33%)	11 (3.67%)
OverallDevelopment	Normal Development: 138 (46%)	Suspected Delay: 162 (54%)	
GrossMotor	Normal Development: 13 (4.33%)	Suspected Delay: 287 (95.67%)	
FineMotorAdaptive	Normal Development: 55 (18.33%)	Suspected Delay: 245 (81.67%)	
ExpressiveLanguage	Normal Development: 91 (30.33%)	Suspected Delay: 209 (69.67%)	
ReceptiveLanguage	Normal Development: 66 (22%)	Suspected Delay: 234 (78%)	
PersonalSocial	Normal Development: 30 (10%)	Suspected Delay: 270 (90%)	

Table 2: Descriptive Statistics of Numeric Variables

Variable	mean	min	Q1	median	Q3	max	Unknown/ Missing
No	150.50	1.00	75.75	150.50	225.25	300.00	0
AgeMother	29.00	16.00	25.00	29.00	32.00	48.00	8
AgeFather	31.80	16.00	27.00	31.00	35.00	59.00	38
AgeCaregiver	38.57	15.00	28.00	33.00	52.00	88.00	0
DiseaseChildWY	1.34	1.00	1.00	1.00	1.75	4.00	218
ElectronicStart	1.71	0.33	1.00	1.67	2.00	4.00	165
DentalCaries1	3.54	1.00	2.00	2.00	5.75	12.00	254
DentalCaries2	2.69	1.00	2.00	2.00	4.00	6.00	254

Table 3: Descriptive Statistics of Categorical Variables

Variable	Value: N (%)		
Province	Chonburi: 300 (100%)		
District	Mueang Chonburi: 100 (33.33%)	Bang Lamung: 199 (66.33%)	Missing: 1 (0.33%)
EduMother	Primary school: 18 (6%)	Second. school: 113 (37.67%)	Assoc. deg/High vocat: 84 (28%)
	Bachelor's degree: 69 (23%)	higher than Bachelor's degree: 5 (1.67%)	No ed.: 2 (0.67%)
	Missing: 9 (3%)		
OccMother	Business owner: 14 (4.67%)	General laborer: 48 (16%)	Government service: 21 (7%)
	Housekeeper: 14 (4.67%)	Unemployed: 51 (17%)	Company employee: 120 (40%)
	Commerce: 24 (8%)	Missing: 8 (2.67%)	
EduFather	Assoc. deg/High vocat: 83 (27.67%)	Primary school: 13 (4.33%)	No ed.: 3 (1%)
	Bachelor's degree: 58 (19.33%)	higher than Bachelor's degree: 2 (0.67%)	Second. school/Vocat. cert.: 103 (34.33%)
	Missing: 38 (12.67%)		
OccFather	Business owner: 29 (9.67%)	General laborer: 64 (21.33%)	Company employee: 123 (41%)
	Housekeeper: 2 (0.67%)	Unemployed: 7 (2.33%)	Missing: 39 (13%)
	Government service: 15 (5%)	Commerce: 21 (7%)	
WhoCaregiver	Mother: 188 (62.67%)	Paternal grandparents: 28 (9.33%)	Maternal grandparents: 63 (21%)
	Father: 11 (3.67%)	Relatives: 10 (3.33%)	
EduCaregiver	Assoc. deg/High vocat.: 60 (20%)	higher than Bachelor's degree: 4 (1.33%)	Second. school/Vocat. cert.: 93 (31%)
OccCaregiver	Bachelor's degree: 49 (16.33%)	No ed.: 7 (2.33%)	Primary school: 87 (29%)
	Business owner: 13 (4.33%)	Unemployed: 107 (35.67%)	Company employee: 63 (21%)
	Housekeeper: 25 (8.33%)	Government service: 15 (5%)	Commerce: 38 (12.67%)
	General laborer: 39 (13%)		
SufficientIncome	Sufficient w/ savings: 136 (45.33%)	Insufficient: 39 (13%)	Sufficient w/o savings: 125 (41.67%)
MaritalStatus	Separated: 62 (20.67%)	Married: 237 (79%)	Widowed: 1 (0.33%)
AgeAtPregnancy	20 – 35: 252 (84%)	< 20: 28 (9.33%)	> 35 years: 20 (6.67%)
AgeAtANC1	12 weeks+: 86 (28.67%)	>= 12 weeks: 214 (71.33%)	
InfectionMother	HIV: 2 (0.67%)	Syphilis: 5 (1.67%)	Missing: 290 (96.67%)
Hemato critMother1	COVID-19: 3 (1%)		
	Hematocrit.>=33%: 214 (71.33%)	Not recorded: 47 (15.67%)	Hematocrit<33%: 38 (12.67%)
	Missing: 1 (0.33%)		

Table 3: Descriptive Statistics of Categorical Variables (Cont.)

Variable		Value: N (%)	
HematocritMother2	Hematocrit. $\geq$ 33%: 203 (67.67%)	Not recorded: 74 (24.67%)	Hematocrit<33%: 23 (7.67%)
AgeAtScreening	9 months: 60 (20%)	18 months: 60 (20%)	30 months: 60 (20%)
	42 months: 60 (20%)	60 months: 60 (20%)	
BirthPlaceProvince	Rayong: 2 (0.67%)	Bangkok: 6 (2%)	Udon Thani: 2 (0.67%)
	Nakhon Pathom: 3 (1%)	Roi Et: 1 (0.33%)	Chiang Mai: 1 (0.33%)
	Bueng Kan: 1 (0.33%)	Sisaket: 1 (0.33%)	Samut Prakan: 1 (0.33%)
	Chonburi: 269 (89.67%)	Surin: 1 (0.33%)	Lampang: 1 (0.33%)
	Prachin Buri: 1 (0.33%)	Missing: 1 (0.33%)	Sakon Nakhon: 1 (0.33%)
	Nonthaburi: 3 (1%)	Chaiyaphum: 1 (0.33%)	Phayao: 1 (0.33%)
	Nakhon Ratchasima: 2 (0.67%)	Kamphaeng Phet: 1 (0.33%)	
ChildEmotions	Easy: 255 (85%)	Fairly difficult: 41 (13.67%)	Difficult: 4 (1.33%)
DiseaseChildWhat	Allergies: 5 (1.67%)	Asthma: 4 (1.33%)	Missing: 285 (95%)
	Anemia: 5 (1.67%)	G6PD Deficiency: 1 (0.33%)	
SymptomChildWY1	Urinary tract infection: 2 (0.67%)	Accident: 4 (1.33%)	Hand, Foot, and Mouth: 1 (0.33%)
	Diarrhea: 10 (3.33%)	RSV: 2 (0.67%)	Missing: 218 (72.67%)
	High fever: 39 (13%)	COVID-19: 4 (1.33%)	Otitis media: 1 (0.33%)
	Pneumonia: 15 (5%)	Asthma: 4 (1.33%)	
HematocritChild1	Hematocrit $\geq$ 33%: 167 (55.67%)	Hematocrit<33%: 15 (5%)	Missing: 118 (39.33%)
HematocritChild2	Hematocrit $\geq$ 33%: 37 (12.33%)	Hematocrit<33%: 2 (0.67%)	Missing: 261 (87%)
BreastfeedingWhat	Breast milk only: 156 (52%)	Breast milk+water: 8 (2.67%)	Breast milk+formula: 89 (29.67%)
	Breastmilk+formula+water: 13 (4.33%)	Missing: 34 (11.33%)	
FolicTake	Eat every week: 117 (39%)	Eat every day: 18 (6%)	Do not eat: 17 (5.67%)
	Missing: 148 (49.33%)		
WatchTVWeekday	> 1 hour: 24 (8%)	$\frac{1}{2}$ -1 hr: 39 (13%)	< 30 minutes: 102 (34%)
	Missing: 135 (45%)		
ElectEquipHour	> 1 hour: 22 (7.33%)	$\frac{1}{2}$ -1 hr: 28 (9.33%)	< 30 minutes: 91 (30.33%)
	Missing: 159 (53%)		
WhenGetDSPM	Received during vaccination: 41 (13.67%)	During prenatal care: 4 (1.33%)	At postpartum ward: 163 (54.33%)
	at home: 7 (2.33%)	Missing: 85 (28.33%)	
StaffTeachUnderstand	Do not understand: 41 (13.67%)	Understand: 139 (46.33%)	Missing: 120 (40%)
ValTeachUnderstand	Do not understand: 8 (2.67%)	Understand: 33 (11%)	Missing: 259 (86.33%)
Playing	Used often: 35 (11.67%)	Used occasionally: 94 (31.33%)	Never: 27 (9%)
	Missing: 144 (48%)		
Use	Never read, never used: 61 (20.33%)	Read, used: 117 (39%)	Read, never used: 37 (12.33%)
	Missing: 85 (28.33%)		
Satisfaction	Most: 21 (7%)	Little/Low: 72 (24%)	Least/Minimal: 1 (0.33%)
	Much: 121 (40.33%)	Missing: 85 (28.33%)	
WtforAge	Fairly underweight: 23 (7.67%)	Fairly overweight: 17 (5.67%)	Overweight: 10 (3.33%)
	Underweight: 7 (2.33%)	Within criteria: 243 (81%)	

Table 3: Descriptive Statistics of Categorical Variables (Cont.)

Variable		Value: N (%)	
Ht for Age	Tall: 5 (1.67%)	Within criteria: 263 (87.67%)	Fairly tall: 4 (1.33%)
	Short: 11 (3.67%)	Fairly short: 17 (5.67%)	
Weight for Height	Starting to be obese: 13 (4.33%)	Plump/Overweight: 9 (3%)	Fairly thin: 16 (5.33%)
	Obese: 6 (2%)	Thin: 13 (4.33%)	Proportional/Normal: 243 (81%)
Oral Health	No dental caries: 202 (67.33%)	Have dental caries: 87 (29%)	No teeth yet: 11 (3.67%)

Instrumentation and Outcome Measures

The primary assessment was conducted using the Developmental Surveillance and Promotion Manual (DSPM), the clinical standard utilized by the Thai Ministry of Public Health (Ministry of Public Health, 2020). The assessment evaluated five distinct developmental domains. 1) Gross Motor (GM) covers the evaluation of postural control and large-muscle coordination. 2) Fine Motor (FM) covers precision and functional use of small muscle groups. 3) Receptive Language (RL) covers linguistic comprehension and processing. 4) Expressive Language (EL) covers communication of thoughts and emotions. 5) Personal and Social (PS) covers self-care and interpersonal navigation. An "Overall Development" score was also calculated to serve as the aggregate dependent measure for holistic developmental health.

Data Preprocessing and Feature Engineering

A total of 66 potential predictors and six developmental dependent variables were analyzed. To ensure data integrity and model readiness, the following preprocessing pipeline was implemented. Initial data auditing identified instances of logical inconsistency. For example, cases where a participant reported "never receiving the DSPM" but subsequently provided details regarding "receiving the manual at the hospital" were corrected to reflect manual receipt. Similarly, conflicting responses regarding "staff explanation" and "understanding the explanation" were harmonized based on the most affirmative response. These corrections ensured that the clinical pathways represented in the data were internally consistent.

Feature Filtering and Feature Engineering

Non-predictive identifiers, such as Sample ID, and

constant variables, such as Province, were removed. A new temporal feature, YearSinceElectronics, was engineered to capture the duration of digital exposure. This was calculated as the difference between the child's current age and the age of initial electronic media use (ElectronicStart). Five numeric predictors were discretized to address non-linear effects and manage sparsity caused by missing values. Parental ages (AgeMother, AgeFather, AgeCaretaker) were binned into 10 developmentally relevant categories ranging from "Early Adolescence" (<15 years) to "Older Adulthood" (≥ 60 years). ElectronicStart and YearSinceElectronics were categorized into seven levels, including "Never" and annual increments up to "4+ years."To preserve sample size while accounting for conditional logic, 10 pairs of "mask-value" features (e.g., Infection status paired with Type of Infection) were merged into single categorical variables. Missing values in dependent sub-features were assigned a "Not Applicable" or "Missing" label based on the primary mask variable. For clinical variables like DentalCaries, missing values were imputed as zero, representing the absence of the condition or the absence of teeth

Mask column merging

To preserve sample size while accounting for conditional logic, 10 pairs of "mask-value" features (e.g., Infection status paired with Type of Infection) were merged into single categorical variables. Missing values in dependent sub-features were assigned a "Not Applicable" or "Missing" label based on the primary mask variable. For clinical variables like DentalCaries, missing values were imputed as zero, representing the absence of the condition or the absence of teeth.

Standardization

Continuous variables were standardized to a mean of 0 and a variance of 1. This is a prerequisite for LASSO-based models, as the algorithm penalizes coefficients based on their magnitude; without standardization, variables with larger scales would be unfairly penalized.

One-Hot Encoding

All categorical variables were transformed via One-Hot Encoding. While this approach typically introduces multicollinearity (the "Dummy Variable Trap"), the use of shrinkage-based models (LASSO) effectively manages collinearity through coefficient penalization. This process expanded the initial feature set from 66 raw predictors to 243 binary and standardized variables for the selection phase. The resulting dataset is very high-dimensional, meaning the number of predictors is relatively large compared to the sample size ($p=243, n=300$). As a result, classical models like linear regression often suffer from overfitting. Overfitting occurs when a statistical model is so complex that it starts to "memorize" the random noise and specific idiosyncrasies of the training data rather than the true underlying trends (Russel & Norvig, 2021). Consequently, an overfitted model may show high accuracy on the current dataset but will fail to generalize to the broader population. This poses a significant risk in nursing research, as it can lead to clinical recommendations based on "fluke" correlations rather than reproducible evidence.

Stability Selection with LassoCV

To resolve the risk of overfitting and ensure the identification of clinically reproducible factors, this study utilized Stability Selection to first reduce the dimensionality of the dataset (Meinshausen & Bühlmann, 2010). Stability selection combines subsampling with a base selection algorithm. Stability selection performs the following steps to select stable predictors. First, the dataset was randomly subsampled. In this study, 50% of the original dataset was subsampled without replacement. Then, a base predictor selection algorithm was applied to the subsample. In this study, Least Absolute Shrinkage and Selection Operator (LASSO) was used. LASSO is a regression equation with a penalty

term added to its loss function as regularization:

$$\left\{ \frac{1}{2n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\}$$

This penalty "shrinks" the coefficients of irrelevant or redundant factors to exactly zero, effectively performing automated feature selection (Ghareeb, 2023; Saputro et al., 2025). Specifically, this study used LassoCV, which was an implementation of LASSO that identifies the optimal alpha constant through internal cross-validation, ensuring the model was tuned for maximum predictive accuracy. This process of subsampling and applying the selection algorithm was repeated for 100 iterations, and the number of times each predictor was selected were recorded. By repeatedly, usually 100 times, subsampling the dataset and applying a selection algorithm on the subsample, only stable predictors will be able to consistently emerge. Predictors that fit well to particular samples but could not generalize (i.e. overfit) would only emerge rarely, as not all samples were included in each repetition. Only "stable" predictors were selected as the final set of predictors. A factor was deemed "stable" if it was selected more than a specified threshold $\pi_{\text{threshold}}$. In this study, a threshold of 0.5 (i.e. predictor must be selected by at least 50% of the repetition) was used to explore plausible predictors. This "consensus" approach ensures that the identified factors are not artifacts of a single random sample but are robust predictors present across many different slices of the data. The process was repeated for each target variable

Confirmatory Analysis using Bootstrapping and Multivariate Logistic Regression

Following the identification of the stable feature set, the magnitude of impact was estimated for each predictor using Odds Ratios (OR) derived from a multivariate logistic regression model. To account for sampling uncertainty and provide robust estimates of effect size, a non-parametric bootstrap approach was utilized with 1,000 iterations. For each iteration, a bootstrap sample ($n=300$) was drawn with replacement from the original dataset. A logistic regression model, restricted to the pre-selected stable features, was fitted to each resampled dataset. In instances of quasi-complete separation or non-convergence (attributable to sparse cells in specific

bootstrap draws), the iteration was excluded from the final distribution to prevent the inflation of parameters. Rather than relying on traditional Wald-based p-values, which can be over-optimistic in post-selection contexts, 95% Confidence Intervals (CI) were derived using the Percentile Method. The Median OR across the successful bootstrap iterations was reported as the primary point estimate. Statistical significance was determined by whether the 95% Bootstrap CI encompassed the null value (OR=1.0).

Results

Feature Stability and Selection Frequencies

Of the 243 initial factors, 13 factors achieved a selection frequency above $\pi_{\text{threshold}}=0.5$ for at least one dependent variable. The distribution of these frequencies for factors is detailed in Error! Reference source not found.. Factors achieving frequencies near 1.0 (100% selection) represent the most reliable "sentinel" indicators for child development in the Chonburi region.

Table 4: Selection Frequencies of Potential Predictors Across Dependent Variables. Variables that were not selected at least 40% in at least one dependent variable were omitted

Factor	Value	OD	GM	FM	RL	EL	PS
DentalPlaque	Have dental plaque	100%*	0%	12%	78%*	91%*	12%
ElectronicStart	2.0-2.9	93%*	0%	46%	1%	7%	46%
PlayingWithChild	None	84%*	0%	27%	7%	36%	23%
Read	Rarely read DSPM	77%*	0%	24%	7%	23%	0%
Read	Never read DSPM	32%	0%	8%	76%*	23%	5%
ElectronicStart	Never	73%*	0%	9%	1%	30%	0%
WhenGetDSPM	When the child received vaccine	70%*	0%	5%	1%	5%	6%
StaffTeach	Understand	64%*	0%	8%	9%	23%	0%
ElectEquip	None	62%*	0%	6%	0%	60%*	0%
HematocritMother2	Hematocrit<33%	59%*	0%	0%	6%	24%	0%
Use	Received, read, and utilized DSPM	51%*	0%	5%	21%	58%*	1%
YearSinceElectronics	Never	53%*	0%	8%	1%	26%	0%
DiseaseChildWY1	Pneumonia	50%*	1%	3%	20%	27%	1%
YearSinceElectronics	1.0-1.9	39%	1%	49%	1%	1%	6%
ChildEmotions	Easy to care for	47%	0%	27%	0%	30%	32%
DiseaseChildWY1	Hand, Foot, and Mouth	47%	0%	30%	0%	0%	0%
WhenGetDSPM	At the postpartum ward	45%	0%	9%	0%	25%	0%
ReadingBook	Never	24%	0%	38%	5%	43%	2%
AgeFather	30-34	42%	0%	2%	0%	18%	0%
BirthPlaceProvince	Phayao	41%	0%	1%	33%	0%	0%
WeightforHeight	Fairly thin	3%	0%	4%	1%	0%	41%
BirthPlaceProvince	Nakhon Ratchasima	40%	0%	14%	5%	7%	21%
BirthPlaceProvince	Samut Prakan	40%	0%	0%	34%	35%	0%
WatchingTV	More than 1 hour/day	40%	0%	4%	19%	3%	4%

* = exceed the threshold $\pi_{\text{threshold}}=0.5$

Results of Feature Selection via Stability Selection

The stability selection process, utilizing a conservative 50% selection frequency threshold $\pi_{\text{threshold}}=0.5$, revealed distinct predictive patterns across the various developmental domains. A significant finding of this

analysis was the disparity in predictability between physical and cognitive/holistic domains. Specifically, for Gross Motor, Fine Motor Adaptive, and Personal and Social development, no features from the 243-factor pool reached the 50% cutoff. This suggests

that for this Chonburi cohort, these domains may be influenced by high individual variance or biological factors not captured within the scope of the regional survey.

Predictors of Overall Development

The Overall Development domain yielded the most comprehensive set of stable predictors, identifying 11 distinct factors related to clinical history, the home environment, and caregiver engagement. Clinical and physiological markers included a history of childhood pneumonia and maternal hematocrit levels below 33% during the second trimester. Oral health emerged as a critical sentinel marker, with the presence of dental plaque being a high-frequency predictor. Environmental and behavioral factors within the home also demonstrated high stability. Predictors included a lack of playtime with the child, the occasional use of the development manual, and the specific commencement of electronic media use between 2.0 and 2.9 years (as well as the "never" category for electronics and electronic equipment). Furthermore, factors related to the DSPM framework were highly stable, specifically receiving the manual during vaccination visits, the caregiver's understanding of staff instructions, and the active utilization and reading of the manual.

Predictors of Language Development

The linguistic domains showed more targeted sets of stable predictors, with oral health consistently

appearing as a common factor. For Receptive Language, stability selection identified only two primary predictors: the presence of dental plaque and the complete failure to utilize the DSPM manual. In contrast, Expressive Language was associated with three stable factors. Consistent with the other domains, the presence of dental plaque remained a significant predictor. Additionally, the absence of electronic equipment use and the proactive reading and utilization of the DSPM manual were identified as stable predictors. The recurrence of dental health and manual utilization across both the linguistic and overall models suggests these factors are the most robust "sentinel" indicators for suspected developmental delay in this population.

Analysis of Stable Predictors for Overall Development

The stable predictors identified in the previous stage were further analyzed using bootstrapped logistic regression to quantify their clinical impact on child development. PlayingWithChild = None was omitted from the analysis since there are only 19 participants with that value, all of whom have regular overall development. Table 5 presents the Median Odds Ratios (OR) and 95% Bootstrap Confidence Intervals (CI). Because the dependent variable represents developmental delay, an OR greater than 1.0 indicates an increased risk, while an OR less than 1.0 indicates a protective factor.

Table 5: Selection Frequencies of Potential Predictors Across Dependent Variables. Variables that were not selected at least 40% in at least one dependent variable were omitted

Variable	Value	Median OR	Lower 95% CI	Upper 95% CI
Intercept		1.29	0.64	2.82
HematocritMother2	Hematocrit<33%	0.16	0.01	0.54
DiseaseChildWY1	Pneumonia	5.96	1.45	7.05×10 ¹²
ElectEquip	None	1.02	0.15	6.75
ElectronicStart	2.0-2.9	6.95	1.97	42.74
ElectronicStart	Never	0.54	0.19	1.37
YearSinceElectronics	Never	0.54	0.19	1.37
WhenGetDSPM	When the child received vaccine	5.47	1.55	25.07
StaffTeach	Understand	0.44	0.18	0.99
Read	Rarely read DSPM	0.38	0.11	1.08
Use	Received, read, and utilized DSPM	0.66	0.23	1.81
DentalPlaque	Have dental plaque	5.36	2.70	12.18

The presence of dental plaque emerged as a powerful predictor of suspected developmental delay, increasing the risk by over fivefold (Median OR = 5.36; 95% CI [2.70, 12.18]). Behavioral and clinical history factors also showed significant associations. Specifically, a history of childhood pneumonia was linked to a higher propensity for delay (Median OR = 5.96; 95% CI [1.45, 7.04e+12]). While the Median OR indicates a strong effect, the extremely wide confidence interval reflects the rarity of this event in the resampled cohorts, suggesting a highly specific but numerically sensitive risk factor. The timing of electronic media introduction proved critical; children starting electronics between the ages of 2.0 and 2.9 years were nearly seven times more likely to experience developmental delay (Median OR = 6.95; 95% CI [1.97, 42.74]). Furthermore, the timing of receiving the DSPM manual was a significant factor.

Children whose families received the manual only at the time of vaccination rather than postpartum were associated with a fivefold increase in the odds of delay (Median OR = 5.47; 95% CI [1.55, 25.07]). Conversely, a maternal hematocrit level below 33% during the second trimester was associated with a lower likelihood of suspected delay (Median OR = 0.16; 95% CI [0.01, 0.54]). These results underscore a complex interplay between clinical history, oral hygiene, and the early childhood home environment

Analysis of Linguistic Development Domains

Table 6 and Table 7 detail the bootstrapped impact estimates for Expressive and Receptive Language. The findings indicate that early linguistic outcomes are heavily influenced by digital exposure and active caregiver engagement with developmental tools.

Table 6: The OR of Selected Predictors for Expressive Language

Variable	Value	Median OR	Lower 95% CI	Upper 95% CI
Intercept		0.66	0.39	1.11
ElectEquip	None	0.29	0.16	0.51
Use	Received, read, and utilized DSPM	0.30	0.15	0.53
DentalPlaque	Have dental plaque	3.41	1.87	5.82

Table 7: The OR of Selected Predictors for Receptive Language

Variable	Value	Median OR	Lower 95% CI	Upper 95% CI
Intercept		0.12	0.07	0.19
Read	Never read DSPM	7.04	3.08	16.39
DentalPlaque	Have dental plaque	3.36	1.87	6.40

For Expressive Language (Table 6), the results suggest that avoiding electronic equipment entirely acts as a significant protective factor (Median OR = 0.29; 95% CI [0.16, 0.51]), effectively reducing the odds of delay by approximately 70%. Similarly, proactive engagement with the DSPM manual (specifically receiving, reading, and actively utilizing the tool) was associated with a 70% reduction in the odds of expressive delay (Median OR = 0.30; 95% CI [0.15, 0.53]). In the Receptive Language domain (Table 7), the failure to utilize the DSPM ("Never utilized") emerged as a primary risk factor, increasing the odds of delay sevenfold (Median OR = 7.04; 95%

CI [3.08, 16.39]). Notably, dental plaque remained a consistent and significant risk factor across both domains (Expressive OR = 3.41; Receptive OR = 3.36), reinforcing its role as a stable indicator of the child's broader developmental environment.

Discussion

A primary finding of this study was the robust selection of dental plaque as a stable predictor for both overall and linguistic development. The

consistent association between visible plaque and suspected delay (OR = 5.36) suggests that oral hygiene may serve as a "sentinel marker" for household routines. This aligns with several studies (Abramovitz et al., 2021; Bodrumlu et al., 2018). Poor oral hygiene may reflect broader household challenges that limit language-rich interactions. The analysis revealed a nuanced relationship between digital media and development. The finding that commencing electronic use between ages 2.0–2.9 years significantly increased the odds of delay (OR = 6.95) supports the "displacement hypothesis," where screen time may replace the "serve and return" interactions essential for language acquisition. Conversely, avoiding electronics was protective for Expressive Language (OR = 0.29), aligning with WHO recommendations to delay digital exposure. The timing of clinical engagement with the DSPM manual provided significant insight. Families who only received the manual during vaccination visits showed a fivefold increase in the odds of delay (OR = 5.47). This likely reflects a "reactive" usage pattern, where the manual is sought or provided only after milestones are missed. In contrast, active reading and utilization of the manual were associated with lower odds of expressive delay (OR = 0.30), suggesting that early postpartum distribution and education are necessary to shift from a reactive to a preventive framework. The identification of childhood pneumonia as a stable predictor (OR = 5.96) suggests that early biological stressors and respiratory morbidity may create developmental "detours." While the wide confidence interval indicates the need for larger samples to stabilize this specific estimate, the high Median OR suggests that children with significant respiratory histories require more intensive monitoring by community nurses. The failure of any of the 243 features to stably predict Gross Motor, Fine Motor, or Personal-Social development is a notable result. This suggests that while linguistic and overall trajectories are highly sensitive to the environmental and socioeconomic factors measured, motor skills may follow a more "biologically canalized" maturational schedule that is less susceptible to factors like screen time or parental education in this specific cohort. In conclusion, the results of this two-stage analytical approach, combining Stability Selection with

Bootstrapped Estimation, suggest that pediatric nursing interventions in Chonburi should prioritize parental health literacy and digital hygiene. By focusing on these stable, high-impact predictors, healthcare providers can better target high-risk families, ensuring the DSPM is used as a proactive roadmap rather than a reactive diagnostic tool. Based on these findings, caregivers across home, daycare, and school settings should monitor oral hygiene and electronic device use as early indicators of developmental risk.

Limitation

The data for this study were collected from two specific districts within Chonburi, Thailand, and only from 300 participants. While the Chonburi population reflects a degree of socioeconomic diversity, combined with a relatively low sample size compared to the population, the findings may not be fully representative of the diverse rural, urban, and regional variations found across the entirety of Thailand. Stability selection is a rigorous tool for high-dimensional data, yet its ability to detect subtle patterns is inherently tied to sample size. In this study, the analysis was focused on the main effects of clinical and environmental factors. A larger sample size would permit the investigation of complex interaction terms, such as the potential buffering effect of regular parental play and singing on children with high electronic media usage. The analysis of rare clinical events, such as childhood pneumonia, encountered significant numerical instability during the bootstrapping phase. This sparsity occurs when a rare factor is absent in specific random bootstrap draws, leading to quasi-complete separation or non-convergence. While these factors were selected for their selection stability, the resulting "exploding" coefficients in certain iterations led to exceptionally wide 95% confidence intervals. The failure of the 243 engineered features to stably predict Gross Motor, Fine Motor, or Personal-Social development suggests a potential "measurement gap" in the survey instrument. It is possible that the predictors of these specific domains were not captured within the scope of the 66 initial variables.

Recommendations

Future research should expand data collection beyond the Chonburi region to include diverse urban and rural provinces across Thailand. Such expansion is necessary to validate whether the stable predictors identified here remain consistent across the country's varied socioeconomic landscapes and regional healthcare infrastructures. Subsequent studies utilizing larger cohorts should move beyond the analysis of main effects to explore complex higher-order interactions. Specifically, research should investigate if protective factors, such as high-frequency caregiver play and singing, can effectively buffer or moderate the developmental risks associated with early electronic media exposure. To resolve the numerical instability associated with rare clinical events like childhood pneumonia, future studies should leverage data from large-scale national clinical registries. This would provide the statistical power necessary to generate precise effect size estimates and stable confidence intervals for infrequent but high-impact biological stressors. To address the potential measurement gap in motor and social development, future survey instruments should incorporate specialized variables such as objective nutritional markers, physical activity metrics, and peer interaction quality. Expanding the scope of initial variables may uncover the environmental predictors underlying these currently "null" developmental domains.

Acknowledgement

We extend our sincere gratitude to all the study participants, their families, and public health personnel from the Development of Maternal and Child Health Care Group, Region Health Promotion Center 6 Chonburi, Thailand, Chonburi Provincial Health Office and Boromarajonani College of Nursing, Chonburi for their participation in this study.

Reference

- Abramovitz, I., Zini, A., Atzmoni, M., Kedem, R., Zur, D., Protter, N. E., & Almozino, G. (2021). Cognitive performance and its associations with dental caries: Results from the dental, oral, medical epidemiological (DOME) records-based nationwide study. *Biology*, 10(3), 178.
- Bodrumlu, E. H., Demiriz, L., & Toprak, S. (2018). Relationship between severe early childhood caries and dental development. *European journal of paediatric dentistry*, 19, 156-160.
- Bornstein, M. H., Hahn, C. S., & Suwalsky, J. T. (2013A). Physically developed and exploratory young infants contribute to their own long-term academic achievement. *Psychological Science*, 24(10), 1906-1917. <https://doi.org/10.1177/0956797613482721>
- Bornstein, M. H., Hahn, C. S., & Wolke, D. (2013B). Systems and cascades in cognitive development and academic achievement. *Child Development*, 84(1), 154-162. <https://doi.org/10.1111/j.1467-8624.2012.01849.x>
- Das, A. C., Dai, R., Lokshin, A., Salam, S., & Smith, L. (2026). Identifying Predictive Combinations of Biomarkers for Early Cancer Detection with Stability Selection in Combination with Ensemble Learning. *Statistics in Biosciences*, 1-29.
- Ghareeb, Z. F. (2023). A comparative study between shrinkage methods (ridge-lasso) using simulation. *Periodicals of Engineering and Natural Sciences*, 11(2), 36-47.
- Hyde, R., O'Grady, L., & Green, M. (2022). Stability selection for mixed effect models with large numbers of predictor variables: A simulation study. *Preventive Veterinary Medicine*, 206, 105714. <https://doi.org/10.1016/j.prevetmed.2022.105714>
- Jantanavivat, W., Nopmaneejumrulers, K., & Seree, P. (2022). Effects of the language development program based on early childhood developmental surveillance and promotion. *International Journal of Child Development & Mental Health*, 10(2).
- Maneekunwong, K., Trakooltorwong, P., Kodyee, S., Wuthipan, N., Sopajaree, C., & Moonpanane, K. (2025). Electronic device usage during

- COVID-19 and suspected developmental delay in rural Thai children: A case-control study. *Asia-Pacific Social Science Review*, 25(3), 4.
- Meinshausen, N., & Bühlmann, P. (2010). Stability selection. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(4), 417–473. <https://doi.org/10.1111/j.1467-9868.2010.00740.x>
- Ministry of Public Health. (2020). *Developmental surveillance and promotion manual (DSPM)*. Department of Health.
- Räikkönen, K., Forsen, T., Henriksson, M., Kajantie, E., Heinonen, K., Pesonen, A. K., ... & Eriksson, J. G. (2009). Growth trajectories and intellectual abilities in young adulthood: The Helsinki Birth Cohort study. *American Journal of Epidemiology*, 170(4), 447–455. <https://doi.org/10.1093/aje/kwp147>
- Rithipukdee, N., & Kusol, K. (2022). Factors associated with the suspected delay in the language development of early childhood in southern Thailand. *Children*, 9(5), 662. <https://doi.org/10.3390/children9050662>
- Russell, S. J., & Norvig, P. (2021). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson.
- Saputro, D. R. S., Wahyu, N. L., & Widyaningsih, Y. (2025). Performance of ridge regression, least absolute shrinkage and selection operator, and elastic net in overcoming multicollinearity. *Journal of Multidisciplinary Applied Natural Science*, 5(2), 370-382.
- Schneider, W., Niklas, F., & Schmiedeler, S. (2014). Intellectual development from early childhood to early adulthood: The impact of early IQ differences on stability and change over time. *Learning and Individual Differences*, 32, 156–162. <https://doi.org/10.1016/j.lindif.2014.02.001>
- Thaineua, V., Sirithongthaworn, S., Kanshana, S., Isaranurak, S., Karnkawinpong, O., Benjaponpitak, A., ... & Tangviriyapaiboon, D. (2024). A 9-year retrospective cohort study of the monitoring and screening of childhood developmental delay in Thailand. *Child: Care, Health and Development*, 50(2), e13233. <https://doi.org/10.1111/cch.13233>
- Wehrens, R., Franceschi, P., Vrhovsek, U., & Mattivi, F. (2011). Stability-based biomarker selection. *Analytica Chimica Acta*, 705(1-2), 15–23. <https://doi.org/10.1016/j.aca.2011.02.025>
- Zhou, J., Sun, J., Liu, Y., Hu, J., & Ye, J. (2013). Patient risk prediction model via top-k stability selection. *In Proceedings of the 2013 SIAM International Conference on Data Mining* (pp. 55–63). Society for Industrial and Applied Mathematics. <https://doi.org/10.1137/1.9781611973310.7>
- Zhu, Z., Chang, S., Cheng, Y., Qi, Q., Li, S., Elhoumed, M., ... & Sudfeld, C. R. (2019). Early life cognitive development trajectories and intelligence quotient in middle childhood and early adolescence in rural western China. *Scientific Reports*, 9(1), 18315. <https://doi.org/10.1038/s41598-019-54848-x>