Machine Learning to Support Visual Inspection of Data: A Clinical Application

Tessa Taylor[1,2] and Marc J. Lanovaz[3]

[1]University of Canterbury/Te Whare Wānanga o Waitaha

[2]Paediatric Feeding International

[3]Université de Montréal

Corresponding author: Dr. Tessa Taylor, University of Canterbury/Te Whare Wānanga o Waitaha, Private Bag 4800, Christchurch 8140, New Zealand.

DrTaylor@PaediatricFeedingIntl.com

Abstract

Practitioners in paediatric feeding programmes often rely on single-case experimental designs and visual inspection to make treatment decisions (e.g., whether to change or keep a treatment in place). However, researchers have shown that this practice remains subjective, and there is no consensus yet on the best approach to support visual inspection results. To address this issue, we present the first application of a paediatric feeding treatment evaluation using machine learning to analyse treatment effects. A 5-year-old male with autism spectrum disorder participated in a 2-week home-based, behaviour-analytic treatment programme. We compared interrater agreement between machine learning and expert visual analysts on the effects of a paediatric feeding treatment within a modified reversal design. Both the visual analyst and the machine learning model generally agreed about the effectiveness of the treatment while overall agreement remained high. Overall, the results suggest that machine learning may provide additional support for the analysis of single-case experimental designs implemented in paediatric feeding treatment evaluations.

Keywords: Artificial Intelligence; Interrater Agreement; Machine Learning; Redistribution; Visual Inspection

Machine Learning to Support Visual Inspection of Data: A Clinical Application

Practitioners and researchers in behaviour analysis have adopted single-case experimental designs to make rigorous treatment decisions. Single-case designs involve the systematic introduction (and in some designs withdrawal) of treatment to examine its effects on a dependent variable (e.g., behaviour). The main method of analysis for single-case designs remains visual inspection, which involves visually examining characteristics of the data such as level, trend, stability, immediacy, overlap, and consistency (Kratochwill et al., 2010; Lane & Gast, 2014). Interpretation of single-case designs is highly contextual and requires more knowledge and skill above objectively analysing the quantitative data. Determining if a change in behaviour is meaningful or not depends on the individual circumstances, situation, and targets. Visual inspection provides numerous benefits including serving as a conservative measure to identify clinically significant behavioural changes. Similar issues have been discussed in other areas of psychology, for example, clinical versus statistical significance. Thus, visual inspection remains subjective in nature.

Consistent with the subjective quality of visual inspection as a methodology, results of research on interrater reliability have been inconsistent (DeProspero & Cohen, 1979; Ninci et al., 2015; Wolfe et al., 2016; Wolfe et al., 2018). More investigation is needed on the reliability of visual inspection and the factors that impact reliability, such as type of data (simulated, research, or clinical), context details available, rating scale type (dichotomous, continuous, number of anchors), number of data points in phases, rater expertise (clinicians or researchers, years of experience), scaling of the y-axis in the phase comparisons, and instructions and definitions provided (Ford et al., 2020; Kahng et al., 2010; Ninci et al., 2015). Unpublished data may have more instability or fewer data points. Previous research on the interrater reliability of visual inspection has examined the impact of clinical context factors, but with only variables desired to increase, on a consistent vertical scale to 100, with a minimum of 5 data points per phase, and from published research datasets (Ford et al., 2020). Therefore, additional research on reliability is needed.

Beyond reliability, results of research on the validity (accuracy and error rates) of visual inspection have also yielded mixed results (DeProspero & Cohen, 1979; Fisher et al., 2003; Kahng et al., 2010). Researchers need to further develop objective methods to add to clinical visual inspection practices and decision-making. In addition to improving reliability of visual inspection, objective methods could help standardise interpretation methodology and improve communicability and relatability of results. Objective methods could also provide quantitative data on the quality, size, and characteristics of the treatment effect. Most importantly, improving analysis methods could benefit the clients served by decreasing decision-making errors about treatment results and effectiveness. As interpretation of single-case designs still requires clinical context and subjective decision making, these objective methods have been proposed to supplement and assist with, rather than replace, visual inspection, and improve reliability and validity.

To address these concerns, researchers have developed a variety of approaches to support visual inspection such as structured criteria and quantitative analyses. One approach involves structured visual inspection criteria with descriptive lines (e.g., median, slope, confidence bands) as guidance (Fisher et al., 2003; Hagopian et al., 1997; Kratochwill et al., 2013; Manolov & Vannest, 2019). Another approach includes statistical quantitative methods such as analysis of overlap of data points between phases, regression-based procedures, interrupted time series analysis, and effect size adaptations from group statistics (see Manolov & Moeyaert, 2017). Despite these various options, visual inspection currently remains the primary, and often sole, method used by many in research and practice.

A practitioner in real world clinical settings may not have extensive training and experience in structured visual inspection criteria and statistical analyses. The applicability of these approaches varies depending on data and design characteristics such as number of data points in each phase, baseline trend, immediacy of effects, variability, trends, and floor and ceiling effects. Thus, knowledge is required to decide which approach to use given the data, and also on how to interpret the results. Practitioners may not have time or knowledge to perform an extensive literature review on these procedures and learn how to implement and interpret them, and may become overwhelmed by the numerous options and lack of consensus in the field.

A novel, different approach is to use artificial intelligence in the form of machine learning to support visual inspection results (Lanovaz et al., 2020). Machine learning has been adopted in other fields such as medicine and education to help in decision-making (Korkmaz & Correia, 2019; Rajkomar et al., 2019). This approach involves a trained computer model that has “learned” to make visual inspection. Machine learning may consider all six data aspects considered in visual inspection including both within-phase and between-phase characteristics. The conceptual appeal of this novel approach is the ability to take valuable, subjective human expertise, clinical experience and practice; objectify and quantify it; and transfer it to be stored and used easily via a trained computer model that has had experience and practice from thousands of data sets. Even if it did not serve as the primary mode of analysis, this method could provide an objective and quantitative second opinion or interrater agreement for support and validation of subjective visual inspection interpretation without the time of another person. It could also be used to support efficiency and objectivity in providing training to practitioners in visual inspection.

Compared to a structured visual aid method (dual-criteria; Fisher et al., 2003), Lanovaz et al. (2020) found machine learning models had higher correspondence with visual interpretation, as well as higher accuracy and power (true positives) with less false positive errors. In this report, we used the stochastic gradient descent (SGD) classifiers model. This model has been shown to have the best properties and functioning for this type of analysis thus far, with higher accuracy and power (Lanovaz et al., 2020). Stochastic gradient descent works in a similar fashion to behavioural shaping. The model makes predictions and then updates itself to produce closer approximations (i.e., less error) during each iteration. That is, it runs successive iterations, keeps the most accurate ones, applies corrections, and keeps running until error is minimised and predictions match the true outcome result.

The machine learning models developed by Lanovaz et al. (2020) to support visual inspection results provide continuous, sensitive results from 0 to 100% in an easily understandable and relatable format in terms of probability of a clear treatment change or effect. The analysis can be easily and quickly performed by a practitioner without any statistical training or software via a simple web app. The web app (https://labrl.shinyapps.io/MachineLearningABGraphs/) can handle a comprehensive range of data characteristics such as clinical data with only 3 data points in a phase, floor and ceiling effects, unstable baseline trends, and delayed/progressive effects. These characteristics make this approach more universally applicable and eliminate decision-making on which of the many quantitative options to use based on the dataset characteristics.

It is important to evaluate methodology to support visual inspection with actual clinical data in context, as well as provide examples for practitioners on how to apply it in real world settings. In the area of behaviour-analytic treatment for paediatric feeding disorders, several clinical characteristics of the data can make interpretation complex and serve as a useful evaluation for objective visual inspection support methods such as machine learning. Consumption, or swallowing, is often the primary variable decisions are based on, while also considering multiple additional variables such as inappropriate mealtime behaviour and negative vocalisations. In a clinical environment, treatment changes may be made more rapidly than in a research setting as resources may be more limited. If a child reaches a time cap without consuming all programmed bites due to packing (holding food in the mouth without swallowing; high latency to mouth clean), a treatment decision may be made after only one or two sessions rather than continuing the condition to optimal data and stability. As another example, if consumption is high and stable at 100%, condition changes may be made primarily based on consumption, while other variables (e.g., negative vocalisations) may not be optimal and stable. A child may accept certain foods or textures causing instability in the data. For example, a child may accept and swallow one pureed food in a few seconds, refuse to accept others, and swallow a higher texture food in 40 s, causing instability in latencies.

Other variables may have an impact across time as different treatment components are added (Sevin et al., 2002), resulting in ineffective treatment phases before the final effective treatment phase. Extinction bursts may occur with addition of additional treatment components, and there may be delayed or progressive effects due to these bursts. Expulsions (spitting out) and packing (not swallowing) may remain low in baseline before escape extinction procedures are added because no food was accepted into the mouth. Multiple ineffective phases may also be present in paediatric feeding treatment evaluations due to least restrictive practice, such as trialling differential reinforcement for attention or tangibles alone prior to escape extinction procedures.

Another benefit to applying machine learning to paediatric feeding data compared to other options is that machine learning is not highly dependent on baseline trend for predictions. In paediatric feeding, we may observe a highly stable baseline with only 3 sessions and ceiling/floor effects for variables such as mouth clean percentage (consumption), latency to acceptance, and inappropriate mealtime behaviour. Additionally, baseline replication consistency may be lower due to graduated exposure and skill development, and delayed effects may occur.

Paediatric feeding also involves targeted clinical cut-offs and goals to consider, such as with latency to swallowing and negative vocalisations, that may impact machine learning results. For latency to swallowing, the aim is to decrease, but the target range is not to reach zero, but rather, within 30 to 45 s depending on food texture, bolus size, and the child's chewing skills. For negative vocalisations, the scale is from 0 to 100%, and the clinical goal may be below 20% of session duration. These characteristics may impact machine learning results because the decisions are made for each AB comparison and are not normalised across all of the phases in the entire treatment evaluation. For both of these variables, an AB comparison normalised for two phases with low levels (e.g., latencies under 60 s compared to reaching a timecap at 600 s; percentages under 25% compared to those over 75%) may appear as a clear change where they would not in the context of a full treatment evaluations on the same maximum vertical axis scale. As another example, dangerous behaviour such as pica where one instance can be significant risk of harm require zero or near zero levels for treatment to be deemed effective. However, these issues are not unique to machine learning, but present in visual inspection with blinded context and other objective methodologies (e.g., structured criteria, statistics) that have been developed.

The purpose of this report was to present a real clinical case application using machine learning to support visual inspection of treatment results. We compared results from machine learning to visual inspection for a paediatric feeding treatment evaluation. We evaluated machine learning as an objective supplement to subjective clinical visual inspection, because it has been found to have higher reliability and validity, providing more consistent results and lower error rates.

Method

Participant, Setting, and Materials

Julien was a 5-year-old Australian and Greek male with autism spectrum disorder (level 2), severe constipation, laxative dependence, recent history (5 months prior) of baby bottle and formula dependence, liquid refusal, medication refusal, viral asthma, and eczema. He did not eat any fruits or vegetables. For proteins, he only ate brand and flavour specific pouch yogurt or custard and red sausages. For starches, he ate any shape of pasta with butter and cheese, rice plain or with some teriyaki sauce (very occasionally with a small amount of teriyaki chicken or salmon next to it), hot dog bun, hot chips, and 2-minute noodles (preferred raw). The participant would eat spaghetti bolognese, but would detect if vegetables were put in it and not eat them. Before treatment, he ate nine foods in total. Julien did not drink enough fluid or eat enough volume consistently. At school, he did not drink water all day. He would spit out certain medications. Julien had the ability to use a fork and spoon, but used his hands instead. He did not sit for meals or eat at the table.

Julien took two bowel management supplements daily as well as laxatives as needed. He had had impactions and bowel stretching with decreased sensation that impacted toilet training attempts. He had had many different bowel regimen trials. Upon admission, he had a bowel movement approximately every other day, but withheld stool and required rewards. Julien had received therapy attempts since 18 months of age aiming to improve his eating and drinking including speech therapy, occupational therapy, multidisciplinary autism feeding team, behavioural therapy, psychology, a sensory book, a wide variety of approaches and strategies, and multiple programmes in an autism specific programme. Julien could speak in full sentences (he had just started speaking a year prior). The participant attended a special kindergarten programme. Problem behaviour included biting, headbutting, crying, and elopement. He had sleep problems and was on melatonin.

This case history was conducted in a behaviour-analytic, short-term and intensive in-home feeding programme as described by Taylor et al. (2020) with no other therapies (all other services were paused). A physician filled out a form to clear and approve participation. We conducted sessions in a family home dining room equipped with seating (Rifton® Activity Chair, child's wooden adjustable chair), dining room table and chairs, laptop computers for data collection, webcam, digital scale, timers, preferred tangible items (e.g., toys, games, electronics), and meal-related materials (e.g., adaptive shallow bowl maroon spoons, rubber coated metal infant spoons, small plates and bowls, infant gum brush, absorbent blue and white pads, gloves, smocks, infant finger gum brush guard, handheld infant food masher bowl, cleaning supplies). Family socioeconomic status was high to medium (informally based on income and employment status). The parents consented to their child's data being used for research and this research project was approved by the second author's university research ethics board.

Response Measurement, Interobserver and Interrater Agreement, and Procedural Integrity

A trained observer recorded participant and feeder behaviours live while seated in the room with laptop computers using a specialised real-time data collection programme (BDataPro) for the treatment evaluation (Bullock et al., 2017). We measured variables only while the demand to eat was in place during bite presentation. Table 1 provides operational definitions for the dependent variables. Latency to mouth clean is a continuous and more sensitive measure for packing which has been defined as not swallowing (i.e., mouth clean; opposite of packing) within a certain time (i.e., 30 s; this varies based on food texture, bolus size, and individual chewing skills). For permanent product data, we took pictures of meal plates consumed and weighed grams consumed on a kitchen scale.

Insert Table 1 about here

An independent observer collected data across phases and conditions using videotaped sessions to assess interobserver agreement (reliability) for 34.8% of sessions. Using BDataPro, we separated sessions into 10-s intervals and calculated proportional agreement between the two observers within each interval.

Interobserver agreement for the dependent variables is presented in Table 1, and averaged 99.6% (range 93.3-100%) for the side deposit and 100% for redistribution procedures. We assessed procedural integrity for 100% of sessions. Observers scored *incorrect integrity* using BDataPro when the feeder made errors of commission (e.g., incorrect attention, bite removal, tangible delivery) and omission (e.g., praise, prompting, physical guidance, tangible delivery) within 3 s of the programmed procedure. The rate of incorrect procedural integrity averaged 0.01 RPM (range 0.1-0.2 RPM). For all sessions scored by a second observer (34.8%), the second observer also assessed interobserver agreement for procedural integrity, which averaged 99.8% (range 96-100%) across all procedural integrity measures.

For clinical visual inspection for each phase comparison, raters responded "yes" or "no" to this question: "Would the change observed from one phase to the next be indicative of functional control of the behaviour in the planned direction (i.e., increase or decrease) if it were reversed and replicated?" Raters also provided a continuous value from 0 (certainty of no effect) to 10 (certainty of an effect), with 0 to 4 corresponding to "no" and 5 to 10 corresponding to "yes." The first author, who was also the case clinician, provided ratings. For interrater agreement, a second independent doctoral-level behaviour analyst and licensed psychologist provided ratings. This second rater was blind to the case details and made ratings based on only the treatment evaluation graphs depicted in Figures 1 and 2. Interrater agreement for the dichotomous ratings was 94.4% overall, and for the continuous ratings was high (Spearman's rho [108] = .92, p < .001). Interrater agreement for each dependent variable is presented in Table 1.

Experimental Design

We used a modified withdrawal/reversal single-case experimental design (ABCDEFAF) to demonstrate experimental control: A was baseline escape, B was differential attention and contingent access, C was differential attention and contingent and noncontingent access plus nonremoval of the spoon and re-presentation of expulsions, D added finger prompt and side deposit, E added move-on, and F added redistribution.

Procedure

A trained doctoral-level behaviour analyst (the first author) conducted sessions approximately 7 hr per day with a trained assistant present. The feeding programme was 11 days (including generalisation and caregiver training) in total with a break on Saturday and the two final days as half days. We completed the treatment evaluation by Day 3 after which we targeted other goals, generalisation, and caregiver training. Sessions were 4 programmed bites, and no new bites were presented after 10 min. The number of sessions and conditions conducted per day varied based on progress. We ran as many sessions per day as possible and took periodic short breaks as needed between sessions. From Day 1 to 3, Julien participated in 11, 23, and 12 sessions, respectively. Sessions averaged 6.1 min (range 2.2 to 10.9 min). Sessions in the entire treatment evaluation totalled 278.1 min (4.6 hr).

We initially targeted eight foods from all food groups that Julien did not currently eat nominated by parents (see Supplementary Information for the complete food list and sample meal plate pictures). Figure 3 depicts food variety per day. We included some previously and inconsistently consumed foods and variations of consumed foods. We increased number of sessions and variety gradually after high and stable consumption. Bolus size was a level small maroon spoon (approximately 1 gram) and 1 cm square bites. Bolus size and programmed bites per session remained unchanged throughout this treatment evaluation.

We preloaded the spoon and placed it on a small child’s plate presenting bites in a self-feeder format for the initial presentation. We presented bites by placing the preloaded spoon and plate on the tray in front of Julien and verbally prompting him to “Take a bite.” We conducted mouth clean checks every 30 s by verbally prompting Julien to “show me ah.” If Julien did not open his mouth, we modelled an open mouth with the verbal prompt and gently touched his chin with the tip of our index finger. If Julien still did not open, we gently touched a flipped (180 degrees) rubber coated tip metal infant spoon to his lips along with the verbal and model prompts. If Julien was packing when the timecap (maximum session duration) elapsed, the feeder removed the bite from his mouth (using a rubber coated tip metal infant spoon) prior to starting another session.

Programme goals included accepting bites within 10 s, swallowing bites within 45 s, less than 1 per min of inappropriate mealtime behaviour, less than 20% of session with negative vocalisations, increasing food variety to 16 foods (2 from each food group), training parents to feed meals with less than 1 per min incorrect procedural integrity, and generalising the protocol. We also increased consistent liquid volume and solid volume to decrease formula dependence.

Prior to treatment, the child participated in a paired-stimulus edible preference assessment sessions using procedures similar to those described by Fisher and colleagues (1992) and Kunkel and colleagues (2018). Tangibles were a variety of reserved items informed by parents and based on a paired-stimulus tangible preference assessment.

Treatment Conditions. Treatment conditions were conducted similar to those described by Taylor (2020, 2020, September 25).

Baseline escape. If Julien did not accept the bite within 5 s, we picked up the spoon and held it stationary approximately 2 cm from his lips until he opened his mouth and allowed us to deposit the bite, 30 s elapsed, or he engaged in inappropriate mealtime behaviour or expulsion, whereby we removed the bite from the tray and provided escape from bite presentations for 30 s. If a bite was accepted, we conducted a mouth clean check every 30 s until clean mouth or the timecap elapsed.

Contingent access and differential attention. Sessions were identical to the previous phase with the following changes. If Julien accepted the bite or had a clean mouth at anytime, we provided descriptive praise (e.g., “Great job taking your bite!” “Great job swallowing your bite!”) and interaction. Upon clean mouth, we provided 30-s access to highly preferred items. We used tokens in the form of tally marks on a sheet of paper for contingent access (value based on paired-stimulus tangible preference assessment) exchanged at the end of the session.

Nonremoval of the spoon and re-presentation of expulsion. Sessions were identical to the previous phase with the following changes. We no longer provided escape contingent on inappropriate mealtime behaviour and expulsion (e.g., Hoch et al., 1994; Kerwin et al., 1995). If he did not accept the bite within 5 s, we did hand-over-hand (placed our hand over Julien’s hand with the spoon or bite) and guided the spoon to his upper lip. We followed his mouth with the spoon if he turned his head, and we blocked disruptions and mouth covering. We kept the bite at his upper lip until acceptance or the timecap elapsed. If he opened his mouth at anytime, we inserted the spoon with the exception of gags, coughs, yawns, or emesis (vomiting), following which the bite was placed only after those responses ceased. We scooped up expulsions and re-presented the bite immediately, but if the bite could not be re-presented (e.g., simultaneous expulsion and emesis), we re-presented a new fresh bite of the same food.

Finger prompt and side deposit. Sessions were identical to the previous phase with the following changes. If Julien did not accept the bite, after 10 s of presentation (5 s after hand-over-hand), we implemented the finger prompt by inserting our forefinger between the cheek and upper gumline until reaching the back of the mouth and held it stationary until he opened to accept the bite or 5 s elapsed (Borrero et al., 2013). We did not place any additional pressure and it was still possible for Julien to keep his mouth closed (i.e., clinch teeth) and refuse to accept the bite. We used the finger from our free hand that was not holding the spoon. Our outer finger and fingernail faced his inner cheek, and the inside of our finger faced his upper teeth and gums. We wore purple nitrile powder-free latex-free gloves and trimmed fingernails shorter than the fingertip. After 15 s of presentation (5 s after finger prompt), the feeder implemented a side deposit (Rubio et al., 2015). For regular texture bites, we used our fingers to manually deposit the bite into the cheek in the space created with the finger prompt (Taylor, 2020). For lower textures (Rubio et al., 2015), we transferred the bolus to an infant gum brush keeping the utensils as close to the mouth as possible. Simultaneous with the finger prompt, we then used the infant gum brush to roll the bolus of food onto Julien's tongue (if he opened his mouth) or onto the inside of his cheek (if his mouth was not open) by gently shifting the straightened finger (i.e., finger prompt) away from the gumline horizontally to create space for the infant gum brush to be inserted and rolling the infant gum brush towards the cheek and away from the gumline. We re-presented expulsions immediately using the side deposit procedure.

Move-on. Sessions were identical to the previous phase with the following changes. Instead of waiting for a mouth clean to present the next bite, the behaviour analyst moved on to presenting the next bite at 30 s (maximum of 4 bites). However, the feeder waited an additional 30 s before moving on if Julien was chewing at 30 s.

Redistribution. Sessions were identical to the previous phase with the following changes. If Julien was packing at the 30-s mouth clean check, we held an infant gum brush horizontally and perpendicular to the tongue, and redistributed the bolus to the tongue behind midline by rolling the brush down and out with gentle downward pressure (Gulotta et al., 2005; Sevin et al., 2002). If Julien's mouth was not open, the behaviour analyst used a finger prompt, inserting our straight forefinger between the cheek and upper gumline to the back of the teeth until resistance was met (Borrero et al., 2013). If the bite was masticated (i.e., chewed up; Volkert et al., 2014) or lower texture, the feeder used the infant gum brush to scoop it up from the mouth. If the bite was not masticated, we removed the bite from the mouth using a rubber coated tip infant spoon and then mashed the bite prior to redistribution (Taylor, 2020, September 25). For mashable textures such as avocado, banana, potato, pumpkin, and ripe strawberry, we mashed the food directly onto the infant gum brush manually with gloved hands. For other soft food textures, we used a handheld infant food masher and bowl, and if needed, a binding food such as mashable foods, hummus, baked beans, or Greek yogurt. If any food was left in the masher bowl at the end of all sessions, postsession we did single-bite presentations with the remainder in the bowl under the same treatment protocol (not shown).

Caregiver Training and Generalisation. The behaviour analyst conducted intensive caregiver behavioural skills training with both parents and generalisation (cutlery/crockery, seating, settings) as described by Taylor (2018). On Day 3 after the treatment evaluation was completed, the feeder moved to full plate meals, biting off portions of food (rather than cut-up bites), and added in a cup of water and regular child's utensils (spoon and fork). We faded and simplified contingent access to a visual timer and beat the clock procedure. On Day 6, the behaviour analyst started caregiver training, and parents fed all meals from Day 7 to discharge. We gradually faded our hours as caregiver training progressed. We gradually increased solid volume (see Supplementary Information for sample pictures of plates of meals eaten).

Social Validity. At programme end, parents completed a written discharge questionnaire (on a Likert-type scale ranging from 1 to 5) to assess programme satisfaction (23 items similar to Hoch et al., 1994; Table 1) and social acceptability of treatment (16 items similar to Intervention Rating Profile by Martens et al., 1985).

Follow-up. For 2 weeks postdischarge, caregivers rotated through the food list and record mealtime data on a shared electronic spreadsheet (duration, percentage of meal consumed, and overall meal rating from 1 [terrible/worse] to 10 [great/perfect]), take pictures of mealtime plates, and video meals. At 1 year, the caregivers recorded which foods from the discharge food list the Julien ate, if the redistribution was needed, and to answer a 5-point Likert-type question from 1 (worse than pretreatment) to 5 (resolved) from the satisfaction survey (Hoch et al., 1994) on how their child's appropriate consumption of a variety of foods was compared to before the programme.

Analysis

For each phase change, we applied the machine learning model to determine whether procedures produced a clear change or not. Specifically, our analyses involved the application of the model derived from stochastic gradient descent previously described and developed by Lanovaz et al. (2020).

To train their model, Lanovaz et al. used supervised machine learning, a procedure that involves providing the input and output. The input were eight characteristics of the AB graphs (e.g., mean of phases A and B, slope of trend in both phases) and the output was whether the graph showed a clear change or not (based on expert visual analysis). The stochastic gradient descent algorithm used to develop the model was based on a logistic regression. The logistic regression is a common function devised to make predictions on categorical variables. The stochastic gradient descent is an optimization method wherein the values provided as input to the logistic regression are transformed by weights prior to computing to the function. From the predictions produced by the function, the algorithm calculates the error between the predictions and the provided output, and then updates the weights to reduce this error. This process was repeated to improve the predictions in an iterative manner, which is similar to behavioural shaping. The result of this process is a model that can be used to analyse novel data for which we do not have an output.

In the current study, we did not develop any new models. Lanovaz et al. (2020) have already shown that their model produced less error than a common aid to visual analysis. Our procedures simply involved providing the input data to the models developed by Lanovaz et al. (2020) and examining whether the model derived from the machine learning algorithm (i.e., stochastic gradient descent) concluded that the graph showed a clear change (or not). Note that practitioners may use a free online calculator available at: https://labrl.shinyapps.io/MachineLearningABGraphs/ to facilitate their work. A screenshot of the input screen and sample output from the online calculator can be viewed in Supplementary Information. In the webpage, the user enters in the data for Phase A and Phase B separated by commas, and selects either “increase” or “decrease” for the expected direction of change to obtain the results. For each variable for the continuous visual analysis ratings, we used Spearman’s rho to evaluate correlations with machine learning results. For all variables, we calculated mean baseline percentage change as a measure of effect size using the last 3 sessions in the treatment evaluation as described in Taylor et al. (2020).

Results

Correspondence between Machine Learning and Visual Inspection

Table 2 presents results of the machine learning analysis. Overall correspondence of machine learning results with clinical visual inspection dichotomous ratings was 88% (95 out of 108 phase comparisons) for the first rater and 82.4% for the second rater (89 out of 108). Overall correlation with clinical visual inspection continuous ratings also was good with the first rater (Spearman’s rho [108] = .80, *p* < .001) and adequate for the second rater (Spearman’s rho [108] = .73, *p* < .001). Both raters had the same pattern of agreements and disagreements with machine learning except for six instances where the second rater indicated a clear change where both machine learning and the first rater did not. We discuss correspondence results below based on the first rater unless otherwise specified.

Insert Table 2 about here

For consumption, which is the primary target variable of this feeding intervention, agreements with machine learning were high (88.9% for percentage; 83.3% for latency) as were correlations (.89 for percentage; .86 for latency). For the main effect from initial baseline to addition of the redistribution procedure, the probability of a clear effect result via machine learning was very high, above 99%, for both consumption percentage and latency. For latency to acceptance, agreement (94%) and correlation (.85) were also high, and probability of a clear treatment effect was above 99%. Similar positive results were found for inappropriate mealtime behaviour (high agreement [89%] and correlation [.94]), which had a clear treatment effect probability above 99% with the addition of the nonremoval component as would be expected. Except for negative vocalisations, there was 100% agreement for all other variables when focusing on the last 3 sessions of the phases.

Agreement was 100% for all phase comparisons for expulsions, which occurred at low levels and with brief, low level bursts with phase changes adding additional treatment components. However, expulsions had the lowest correlation with the continuous ratings, which were 0 (certainty of no effect) for all phase comparisons except for Phases E to F when redistribution was added, where the rating was still low at 1. For the initial effective treatment phase when the redistribution was added, for negative vocalisations, probability of a clear effect was high above 98% and agreed with visual inspection.

Treatment Evaluation Results and Data Characteristics

In the next paragraphs, we present the detailed intervention results from the treatment evaluation (Figures 1 and 2). When visual inspection and machine learning disagreed, we discuss potential reasons as we move along through discussing the results depicted in the graphs. For all disagreements with machine learning, the first rater had included comments along with the rating. The second rater also included similar comments for expulsion and negative vocalisations not posing an issue in baseline. Except for negative vocalisations, when disagreements occurred, the first rater’s continuous ratings (ranging from 2 to 8) were not at the extremes of certainty.

Insert Figures 1 & 2 about here

Prior to the treatment evaluation, we conducted two edible preference assessments. Julien consumed some bites of five foods in the first assessment that included some previously or inconsistently consumed foods. He consumed some bites of two foods in the second assessment. In the treatment evaluation, for the first two phases (baseline escape; differential attention and contingent access), consumption was low and stable at 0% and latency to mouth clean was high and stable at 600 s. With nonremoval and re-presentation, visual inspection did not indicate a clear change for consumption (Figure 1, upper panel), but machine learning did, albeit at lower probabilities. The pattern of data from this phase (temporary increase in consumption to 100%) is clinically unusual and the data were highly unstable (consumption: *M* = 50%, range 0-100%; latency to mouth clean: *M* = 309 s, range 9-600 s). With the addition of the finger prompt and side deposit procedures, consumption was low and stable (*M* = 11%, range 0-25%) and latency to mouth clean was high and stable (*M* = 482 s, range 64-600 s) except for 1 session. Visual inspection indicated a clear change for latency to acceptance (Figure 1, lower panel) in contrast to machine learning. Delayed effects were present in both phases compared, as well as unusually low latency to acceptance for the first 5 sessions of nonremoval and re-presentation. With the addition of the move-on component, consumption was low and stable at 0% and latency to mouth clean was high and stable (*M* = 585 s, range 578-595 s). With the addition of redistribution, consumption increased to high and stable levels at 100% (*M* = 83%, range 25-100%) and latency to mouth clean decreased to low and stable levels (*M* = 53 s, range 31-148 s). Machine learning and visual inspection agreed 100% for both the move-on phase (no clear change) and the redistribution phase (clear change).

During the reversal to baseline escape, consumption was on a decreasing trend to 0% (*M* = 35%, range 0-75%), latency to mouth clean was on an increasing trend to 600 s (*M* = 188 s, range 50-600 s), and Julien was only consuming certain foods. In contrast to visual inspection, machine learning did not indicate a clear change result for consumption and inappropriate mealtime behaviour (Figure 2, upper panel) in this phase. For consumption, there was a progressive effect and only one data point at 0% and high latency. Similarly, for inappropriate mealtime behaviour (*M* = 2.4 RPM, range 0.3-8.8 RPM), there was only 1 data point at a high level. This reversal could have been continued to obtain further data points to improve quality and experimental control. When the treatment package was reimplemented, consumption increased to high and stable levels at 100% and latency to mouth clean decreased to low and stable levels (*M* = 64 s, range 40-114 s). Visual inspection determined a clear change for latency to mouth clean (Figure 1, lower panel) and inappropriate mealtime behaviour (*M* = 0.2 RPM, range 0-1.1 RPM) in contrast to machine learning, and as described for the reversal, there were delayed effects and only 1 data point at the high levels. In comparing the initial baseline to both treatment package phases with redistribution, machine learning and visual inspection agreed for all variables except negative vocalisations (Figure 2, upper panel).

For negative vocalisations with baseline escape (*M* = 1.8%, range 0-5.3%) and contingent access and differential attention (*M* = 0%), visual inspection did not indicate a clear change, but machine learning did (although at a lower probability of 66.5%). In baseline, the final data point was increasing at 5%, which is clinically still low for a 100% scaled variable and when the clinical goal is below 20%. From this baseline result, two additional disagreements between machine learning and visual inspection occurred in comparison to the final treatment package phases (F1: *M* = 6.8%, range 0-65%; F2: *M* = 2.5%, range 0-18%). This pattern also occurred in the reversal (slight increase to 3.2%; *M* = 0.9%, range 0-3.2%).

Feeding Programme Outcomes and Follow-up

A table summary of programme outcomes for Julien and a figure of the total number of new foods per day are available in Supplementary Information. Julien’s variety was at 15 foods from all food groups on the first day of treatment. Julien started consuming bites in the first session of treatment (less than 10 min) and reached 100% consumption after 19.6 min of treatment (2 sessions). He consumed bites in the first treatment session on the first day of treatment, and on the second day, was at stable 100% consumption in final treatment in 3 hr (192 min) after 23 sessions (5 sessions and 28 min in the final treatment phase with redistribution). Use of both the side deposit and redistribution procedures quickly decreased to zero. After the programme in the afternoon and evening with his family, Julien ate 13 novel foods from all food groups free access without a protocol (blueberry, broccoli, cucumber, strawberry, watermelon, ravioli, raw carrot, lettuce, chicken, banana, burger patty, sushi, boiled egg). Julien's variety was at 35 foods at regular texture by the end of this treatment evaluation (Day 3). In all full plate meals (not shown), consumption of solids and liquids were high and stable at 100% and use of redistribution was zero. Julien's parents never used the redistribution procedure. Julien met 100% of goals and reached a variety of over 60 foods across food groups at regular texture, independence with self-feeding and scooping with utensils, and a full plate portion presentation. Gains were maintained to a year and parents reported high satisfaction and treatment acceptability.

At 1-year follow-up, parents reported Julien's feeding problems were much better than pretreatment (4 out of 5). Redistribution was not being used. Julien was still eating 73% of the foods from the discharge food list across all food groups. Some foods were not offered as frequently, and some foods decreased during COVID-19 pandemic restrictions along with other behavioural changes. Parents reported he would eat a variety of foods free choice without a protocol (e.g., apple, melon, bananas, peas, carrots, cucumber, sausage, homemade biscuits and cake, yoghurt, ham, sometimes vegemite sandwiches, rice, chicken, broccoli, pasta). Parents reported the feeding intervention was extremely successful, that Julien was growing very tall, able to concentrate better at school, and that constipation was easier to manage.

Discussion

Machine learning results corresponded highly overall (88%) with clinical visual inspection from real full treatment evaluation graphs. We examined the pattern of results and performance using an actual clinical case example in paediatric feeding, particularly where disagreements occurred. The data from this treatment evaluation were not perfect, with phases with 3 data points only, a phase with high instability, and a reversal not continued to full clear stability. In visual inspection, the raters commented about these imperfect characteristics of the dataset in instances when ratings disagreed with machine learning results.

Given the limitations of clinical data, machine learning still performed well, particularly for the for the main treatment effect (99%) and for last 3 sessions in the phases. Except for expulsions and negative vocalisations, correlations of the probability values with the continuous ratings were also high overall (*sr* > .80). For expulsions, the rater was highly certain (between 0 and 1) in the continuous ratings, which negatively impacted the correlations.

For negative vocalisations, with the y-axis scaled to the highest value for the entire treatment evaluation (94%) or 100%, this pattern impacts interpretation of results for low values. Even given this issue with machine learning performance for negative vocalisations, agreement and probability were high (above 98%) for the one phase where there was a treatment effect, we observed several near zero probabilities and high agreements in other phase comparisons without a treatment effect, and all of the "false positives" were low in probability values (in the 60s). Thus, two considerations and future directions for machine learning emerged from the current analysis. One is the possible consideration of clinical ranges for latency to swallowing and for percentage variables such as negative vocalisations. Another is consideration of the scale of the y-axis or range and standardisation to the entire treatment evaluation versus each single AB phase comparison.

Adding to the visual inspection reliability literature, we compared interrater agreement on visual inspection with the case clinician who performed this treatment evaluation and a second rater not privy to case details, used both a dichotomous and 10-point rating scale, for variables desired to both increase and decrease, for both the entire phase and the last 3 sessions, and for phases with only 3 data points from a real clinical case in the context of the full treatment evaluation. Regarding interrater reliability, ratings and agreement were slightly higher for the case clinician, who works in paediatric feeding exclusively and has expertise in this area, versus the second rater. The use of a 0 to 10 scale in order to examine correlations to the 0 to 100% probability output from machine learning was useful information to add beyond the dichotomous decisions. The 10-point scale may provide a better balance between sensitivity and reliability compared to previously used scales of 0 or 1, 1 to 4, or 0 to 100. Future research could examine these results in a larger group of participants. Future research could also examine interrater reliability for labelled and scaled full treatment evaluations versus normalised single AB phase comparisons without knowledge of the phases, variables, scale of measurement on the y-axis, or case. It may be that normalised single AB phase comparisons have lower or higher interrater reliability, but higher correspondence with machine learning results.

Our study has some limitations and future directions to consider. Our application of machine learning was post hoc to examine how it could have supported decisions made. Applying machine learning a priori would be most useful. Machine learning indicated treatment effects (although at lower probabilities) for negative vocalisations when levels were clinically low and insignificant. Future studies could examine this issue and if the machine learning procedures could be modified to handle it. Another future direction for machine learning would be able to examine consistency of the replication phase (i.e., initial baseline to reversal [A1 to A2]; initial treatment to replication [F1 to F2]) where no direction of change is expected. Additionally, machine learning research could be expanded to specifically handle other single-case experimental designs such as multielement and changing criterion. Finally, machine learning provides results in terms of probability of a clear effect reflecting quality of experimental control on the whole and support of visual inspection decisions, but it does not provide a typical effect size convention.

In addition to supplementing visual analysis, instructors may eventually use machine learning to train individuals to analyse single-case graphs. For example, trainees could rate nonsimulated graphs and then compare their analyses to those produced by the models derived from machine learning. The main benefit of such an approach would be that the trainee could practice on their own and receive immediate feedback without the need for a trainer to be present. Alternatively, researchers could develop a fully automated web app with simulated graphs using Monte Carlo simulations. By knowing the true values, it would become possible to compare the analysis of the trainee not only with machine learning, but also with the true value generated by the simulation. Importantly, training to analyse single-case graphs should go beyond using a simple online tool and focus on how the interaction between different dimensions of the data (i.e., level, trend, variability, overlap, immediacy) influences analysis.

This clinical case history report also adds to the literature by presenting a replication of Taylor's (2020, September 25) effective use of redistribution with regular bites of food and extend it by following up at 1-year postdischarge. To our knowledge, this redistribution procedure with regular texture food has only been presented once thus far. In addition, the current case history provides an example of two procedures from a small literature base with only a handful of studies (a side deposit physical guidance procedure with regular texture food [Taylor, 2020b] and a move-on to the next bite component for not swallowing), not achieving consumption, and thus, an additional treatment component of redistribution being added to a treatment package. Redistribution increased consumption to 100% and was no longer needed prior to training and generalisation, and gains were maintained to 1 year. As for limitations in the feeding treatment evaluation, the behaviour analyst did not conduct a component analysis to demonstrate the effectiveness of the redistribution component alone as it was evaluated as part of a treatment package.

In conclusion, the reliability and validity of visual inspection has been a long-standing area of investigation. This subjective clinical interpretation is needed for single-case experimental designs, in addition to objective methodologies to supplement this process, improving consistency and accuracy. Although many approaches have been put forth to support visual inspection, they have not been yet been universally or consistently adopted even by applied researchers much less practitioners. Machine learning could serve as a viable objective supplement to subjective clinical visual inspection, providing more consistent results and lower error rates. This report shows a successful collaboration between a quantitative researcher and a clinician (in an in-home setting outside of hospitals or clinics). This serves as an applied real-world example of how machine learning can be easily used to support visual inspection in treatment evaluations by practitioners. The researchers hope that this example helps to clarify visual inspection support options and decrease barriers to their adoption in practice.

Acknowledgements

Thank you to Dr. Jessica Boisjoli for her assistance with visual inspection ratings. This case history was conducted in the context of a treatment evaluation for a clinical case where progress monitoring and data indicated that additional procedures were needed; thus, it is a retrospective report of

existing non-identifiable data. The authors report no conflicts of interest. Participant name was changed to protect confidentiality. This case history was performed in accordance with ethical standards laid out in the 1964 Declaration of Helsinki and its later amendments and in the Canadian Tri-Council Policy Statement: Ethical Conduct for Research Involving Humans. This study was supported in part by a salary award from the Fonds de recherche du Québec – Santé (#269462) to the second author. Caregivers and behaviour analysts provided written informed consent.

References

Borrero, C. S. W., Schlereth, G. J., Rubio, E. K., & Taylor, T. (2013). A comparison of two physical guidance procedures in the treatment of pediatric food refusal. *Behavioral Interventions*, *28*(4), 261-280. https://doi.org/10.1002/bin.1373

Bullock, C. E., Fisher, W. W., & Hagopian, L. P. (2017). Description and validation of a computerized behavioral data program: "BDataPro". *The Behavior Analyst*, *40*(1), 275-285. https://doi.org/10.1007/s40614-016-0079-0

DeProspero, A., & Cohen, S. (1979). Inconsistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *12*(4), 573-579. https://doi.org/10.1901/jaba.1979.12-573

Fisher, W. W., Kelley, M. E., & Lomas, J. E. (2003). Visual aids and structured criteria for improving visual inspection and interpretation of single-case designs. *Journal of Applied Behavior Analysis*, *36*(3), 387-406. https://doi.org/https://doi.org/10.1901/jaba.2003.36-387

Fisher, W. W., Piazza, C. C., Bowman, L. G., Hagopian, L. P., Owens, J. C., & Slevin, I. (1992). A comparison of two approaches for identifying reinforcers for persons with severe and profound disabilities. *Journal of Applied Behavior Analysis*, *25*(2), 491-498. https://doi.org/10.1901/jaba.1992.25-491

Ford, A. L., Rudolph, B. N., Pennington, B., & Byiers, B. J. (2020). An exploration of the interrater agreement of visual analysis with and without context. *Journal of Applied Behavior Analysis*, *53*(1), 572-583. https://doi.org/10.1002/jaba.560

Gulotta, C. S., Piazza, C. C., Patel, M. R., & Layer, S. A. (2005). Using food redistribution to reduce packing in children with severe food refusal. *Journal of Applied Behavior Analysis*, *38*(1), 39-50. https://doi.org/https://doi.org/10.1901/jaba.2005.168-03

Hagopian, L. P., Fisher, W. W., Thompson, R. H., Owen-DeSchryver, J., Iwata, B. A., & Wacker, D. P. (1997). Toward the development of structured criteria for interpretation of functional analysis data. *Journal of Applied Behavior Analysis*, *30*(2), 313-326. https://doi.org/10.1901/jaba.1997.30-313

Hoch, T. A., Babbitt, R. L., Coe, D. A., Krell, D. M., & Hackbert, L. (1994). Contingency contacting: Combining positive reinforcement and escape extinction procedures to treat persistent food refusal. *Behavior Modification*, *18*(1), 106-128. https://doi.org/https://doi.org/10.1177/01454455940181007

Kahng, S. W., Chung, K. M., Gutshall, K., Pitts, S. C., Kao, J., & Girolami, K. (2010). Consistent visual analyses of intrasubject data. *Journal of Applied Behavior Analysis*, *43*(1), 35-45. https://doi.org/https://doi.org/10.1901/jaba.2010.43-35

Kerwin, M. L. E., Ahearn, W. H., Eicher, P. S., & Burd, D. M. (1995). The costs of eating: A behavioral economic analysis of food refusal. *Journal of Applied Behavior Analysis*, *28*(3), 245-260. https://doi.org/10.1901/jaba.1995.28-245

Korkmaz, C., & Correia, A.-P. (2019). A review of research on machine learning in educational technology. *Educational Media International*, *56*(3), 250-267. https://doi.org/https://doi.org/10.1080/09523987.2019.1669875

Kratochwill, T. R., Hitchcock, J., Horner, R., Levin, J. R., Odom, S., Rindskopf, D., & Shadish, W. (2010). Single-case designs technical documentation. *Retrieved from What Works Clearinghouse website: http://ies.ed.gov/ncee/wwc/pdf/wwc_scd.pdf*.

Kratochwill, T. R., Hitchcock, J. H., Horner, R. H., Levin, J. R., Odom, S. L., Rindskopf, D. M., & Shadish, W. R. (2013). Single-case intervention research design standards. *Remedial* *and Special Education*, *34*(1), 26-38. https://doi.org/https://doi.org/10.1177/0741932512452794

Kunkel, K. R., Kozlowski, A. M., Taylor, T., & González, M. L. (2018). Validating a food avoidance assessment for children with food selectivity. *Behavioral Development*, *23*(2), 89-105. https://doi.org/10.1037/bdb0000078

Lane, J. D., & Gast, D. L. (2014). Visual analysis in single case experimental design studies: Brief review and guidelines. *Neuropsychological rehabilitation*, *24*(3-4), 445-463. https://doi.org/https://doi.org/10.1080/09602011.2013.815636

Lanovaz, M. J., Giannakakos, A. R., & Destras, O. (2020). Machine learning to analyze single-case data: A proof of concept. *Perspectives on Behavior Science*, 1-18. https://doi.org/https://doi.org/10.1007/s40614-020-00244-0

Manolov, R., & Moeyaert, M. (2017). Recommendations for choosing single-case data analytical techniques. *Behavior Therapy*, *48*(1), 97-114. https://doi.org/https://doi.org/10.1016/j.beth.2016.04.008

Manolov, R., & Vannest, K. J. (2019). A visual aid and objective rule encompassing the data features of visual analysis. *Behavior Modification*, 1-32. https://doi.org/https://doi.org/10.1177%2F0145445519854323

Martens, B. K., Witt, J. C., Elliott, S. N., & Darveaux, D. X. (1985). Teacher judgments concerning the acceptability of school-based interventions. *Professional Psychology: Research and Practice*, *16*(2), 191-198. https://doi.org/10.1037/0735-7028.16.2.191

Ninci, J., Vannest, K. J., Willson, V., & Zhang, N. (2015). Interrater agreement between visual analysts of single-case data: A meta-analysis. *Behavior Modification*, *39*(4), 510-541. https://doi.org/https://doi.org/10.1177%2F0145445515581327

Rajkomar, A., Dean, J., & Kohane, I. (2019). Machine learning in medicine. *New England Journal of Medicine*, *380*(14), 1347-1358. https://doi.org/DOI: 10.1056/NEJMra1814259

Rubio, E. K., Borrero, C. S. W., & Taylor, T. (2015). Use of a side deposit to increase consumption in children with food refusal. *Behavioral Interventions*, *30*(3), 231-246. https://doi.org/10.1002/bin.1404

Sevin, B. M., Gulotta, C. S., Sierp, B. J., Rosica, L. A., & Miller, L. J. (2002). Analysis of response covariation among multiple topographies of food refusal. *Journal of Applied Behavior Analysis*, *35*(1), 65-68. https://doi.org/10.1901/jaba.2002.35-65

Taylor, T. (2020). Side deposit with regular texture food for clinical cases in-home. *Journal of Pediatric Psychology*, *45*(4), 399-410. https://doi.org/https://doi.org/10.1093/jpepsy/jsaa004

Taylor, T. (2020, September 25). Redistribution with regular texture food for clinical cases in-home. https://doi.org/10.31234/osf.io/tv5mg

Taylor, T., Blampied, N., & Roglić, N. (2020). Consecutive controlled case series demonstrates how parents can be trained to treat paediatric feeding disorders at home. *Acta Paediatrica*. https://doi.org/DOI:10.1111/apa.15372

Volkert, V. M., Peterson, K. M., Zeleny, J. R., & Piazza, C. C. (2014). A clinical protocol to increase chewing and assess mastication in children with feeding disorders. *Behavior Modification*, *38*(5), 705-729. https://doi.org/10.1177/0145445514536575

Wolfe, K., Seaman, M. A., & Drasgow, E. (2016). Interrater Agreement on the Visual Analysis of Individual Tiers and Functional Relations in Multiple Baseline Designs. *Behavior Modification*, *40*(6), 852-873. https://doi.org/10.1177/0145445516644699

Wolfe, K., Seaman, M. A., Drasgow, E., & Sherlock, P. (2018). An evaluation of the agreement between the conservative dual-criterion method and expert visual analysis. *Journal of Applied Behavior Analysis, 51*(2), 345-351. https://doi.org/10.1002/jaba.453

Table 1

Dependent Variables, Interobserver Agreement (IOA) for Behaviour, and Visual Inspection Interrater Agreement (VIRA)

Variable	Definition	Calculation	IOA (range)	IRA	IRA (*sr*)
Mouth Clean (Frequency)	Product measure of swallowing; no food larger than size of pea in mouth at a 30-s check "show me ah" unless absence of food was due to expulsion	Divided by 4 (total number of programmed bites for the session), multiplied by 100; overall measure of consumption of food (% of bites consumed)	95.9% (76.9-100)	100%	.93
Latency to Acceptance (Duration)	Scored by activating/deactivating a timer to count seconds from bite presentation (prompted "take a bite," bite placed within arm's reach) to acceptance (entire bolus, except an amount smaller than a pea, passed plane of lips into mouth for the first time at any time during each bite presentation	Divided by frequency of occurrence; Average seconds per bite	91.1% (68.4-100)	88.9%	.92
Latency to Mouth Clean (Duration)	Scored by activating/deactivating a timer to count seconds from acceptance to mouth clean	Divided by frequency of occurrence; Average seconds per bite	92.9% (57.9-100)	100%	.92
Inappropriate Mealtime Behaviour (IMB; Frequency)	45 degree or further head turn, touch of feeder's arm below elbow, push bite/plate away, or mouth cover during spoon presentation	Divided by session duration within demand; Responses per minute (RPM)	98.5% (98.1-100)	88.9%	.88
Expulsion (Frequency)	Food larger than size of pea passed plane of lips after deposit	Divided by session duration within demand; Responses per minute (RPM)	99.9% (90-100)	88.9%	.75
Negative Vocalisations (Duration)	Crying, screaming, whining, negative statements about the food/meal; 3-s offset (deactivated when ceased for 3 s)	Divided by session duration, multiplied by 100	91.3% (50-100)	100%	.75

Table 2

Machine Learning Stochastic Gradient Descent (SGD) Results in Percentage Probabilities

Phase Comparison	Mouth Clean	Latency to Mouth Clean	Latency to Acceptance	Inappropriate Mealtime Behaviour	Expulsion	Negative Vocalisations
Baseline Escape versus Contingent Access, Differential Attention	30.2%	30.2%	0.4%	0.1% [a]	0.2%	66.5%*
Last 3	30.2%	30.2%	0.4%	0.1%	0.2%	66.5%*
+ Nonremoval, Re-presentation	76.1%*	68.3%*	0.9%	99.5%*	0%	0.2%
Last 3	19.5%	19.5%	0%	98.8%*	0.1%	0.2%
+ Side Deposit	0%	0%	0.1%	0.1%	0.1%	0%
Last 3	1.9%	0.8%	97.9%*	0%	30.2%	0%
+ Move-on	0%	0%	2.4% [a]	0%	0%	0%
Last 3	0.2%	0%	2.4% [a]	0%	0%	0%
+ Redistribution	99.3%*	99.9%*	1.4%	15.4% [a]	1.6% [a]	98.3%*
Last 3	99.6%*	99.5%*	0.7%	55.8%*	16.0% [a]	99.5%*
Baseline Escape versus Redistribution	99.3%*	100%*	99.9%*	83.9%*	0.4%	0.9%
Last 3	99.6%*	99.6%*	99.3%*	66.5%*	30.2%	66.5%*
Redistribution verses Baseline Reversal	17.3%	10.6%	71.9%*	17.3%	0.1%	0.1%
Last 3	98.5%*	69.3%*	87.9%*	59.5%*	30.2%	66.4%*
Redistribution Replication	96.7%*	10.1%	81.8%*	12.0%	0.4%	0.4%
Last 3	99.2%*	84.6%*	98.0%*	79.3%*	30.2%	45.2%
Baseline Reversal versus Redistribution Replication	99.9%*	99.0%*	99.9%*	92.0%*	0.6%	1.2%
Last 3	99.6%*	99.6%*	99.4%*	84.9%*	30.2%	66.5%*
Rater 1 % Agreement	88.9%	83.3%	94.4%	88.9%	100%	72.2%
Rater 1 Spearman's rho (*df* = 108)	.89* *p* < .001	.86* *p* < .001	.85* *p* < .001	.94* *p* < .001	.21 *p* = .408	.55* *p* = .018
Rater 2 % Agreement	88.9%	83.3%	83.3%	72.2%	88.9%	72.2%
Rater 2 Spearman's rho (*df* = 108)	.83* *p* < .001	.79* *p* < .001	.83* *p* < .001	.87* *p* < .001	-.11 *p* = .664	.46 *p* = .057

* Probability of a clear change from machine learning results.

Shaded cells indicate agreement with dichotomous clinical visual inspection ratings with Rater 1.

[a] Rater 2 disagreements with both machine learning and Rater 1.

Figure 1

Percentage Consumption (Top Panel), Average Latency to Acceptance and Mouth Clean in Seconds (Bottom Panel). For Latency to Mouth Clean/Swallowing, Horizontal Line Data Markers Indicate Reaching the Session Timecap with No Bites Accepted or Consumed.

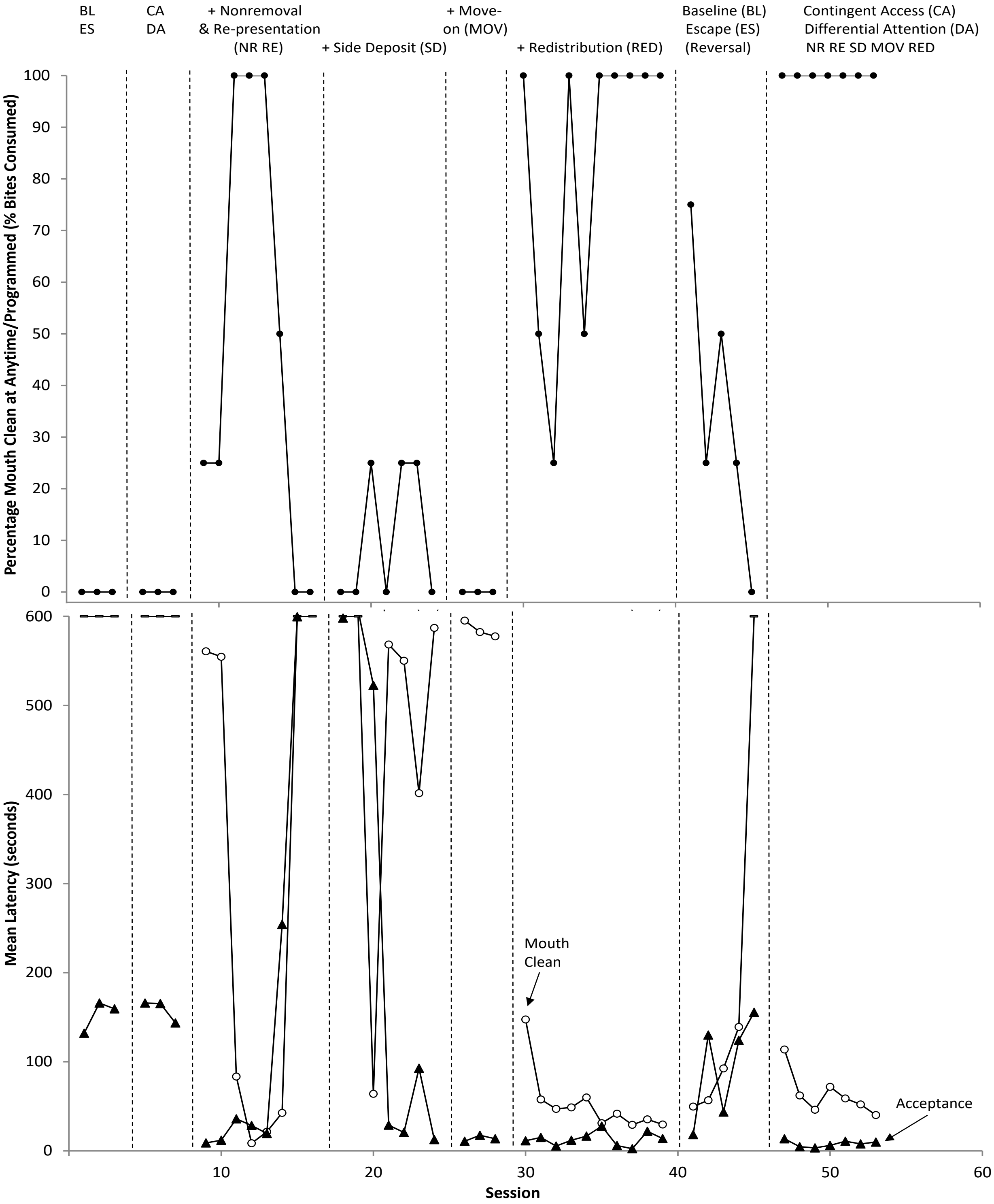
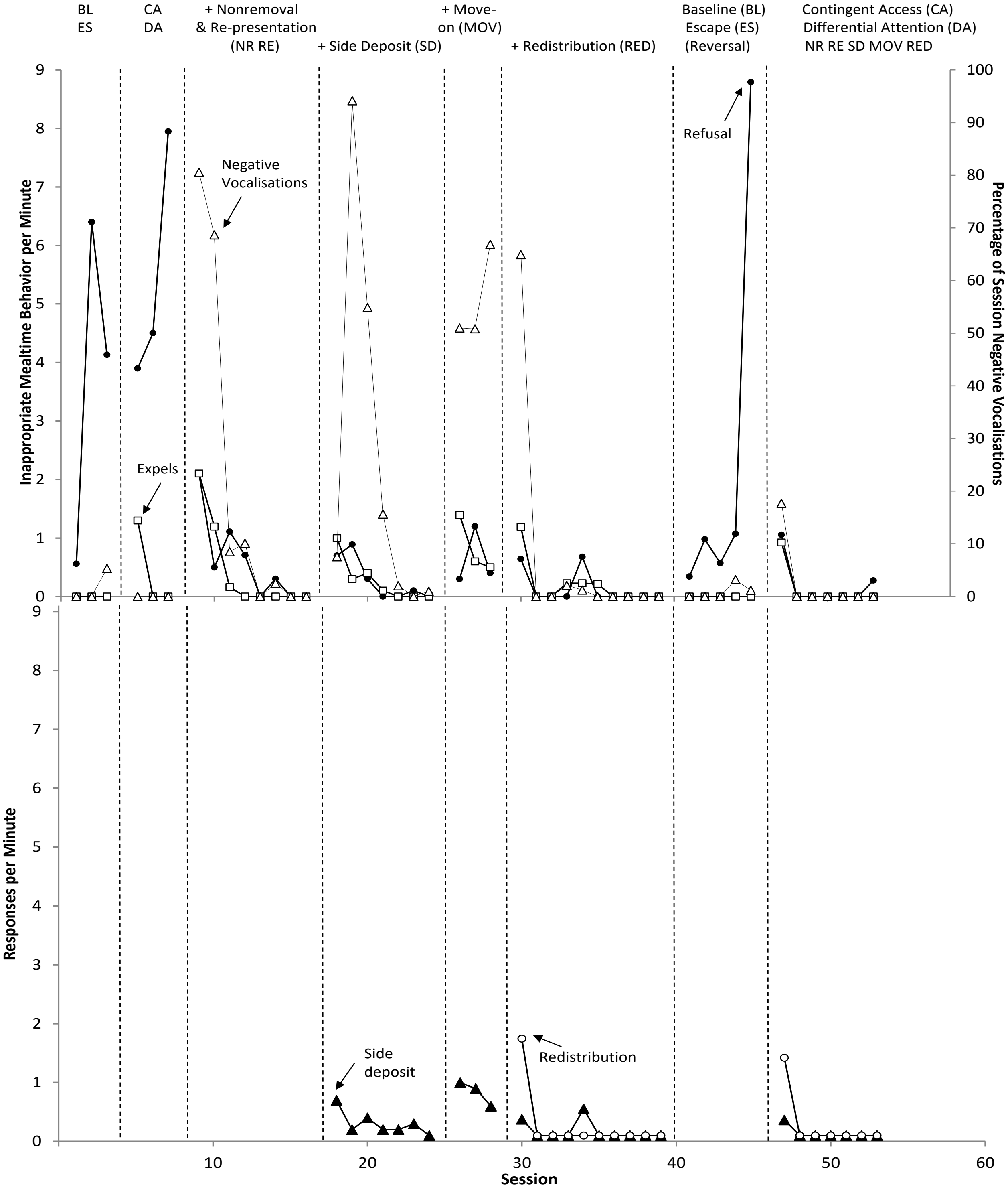


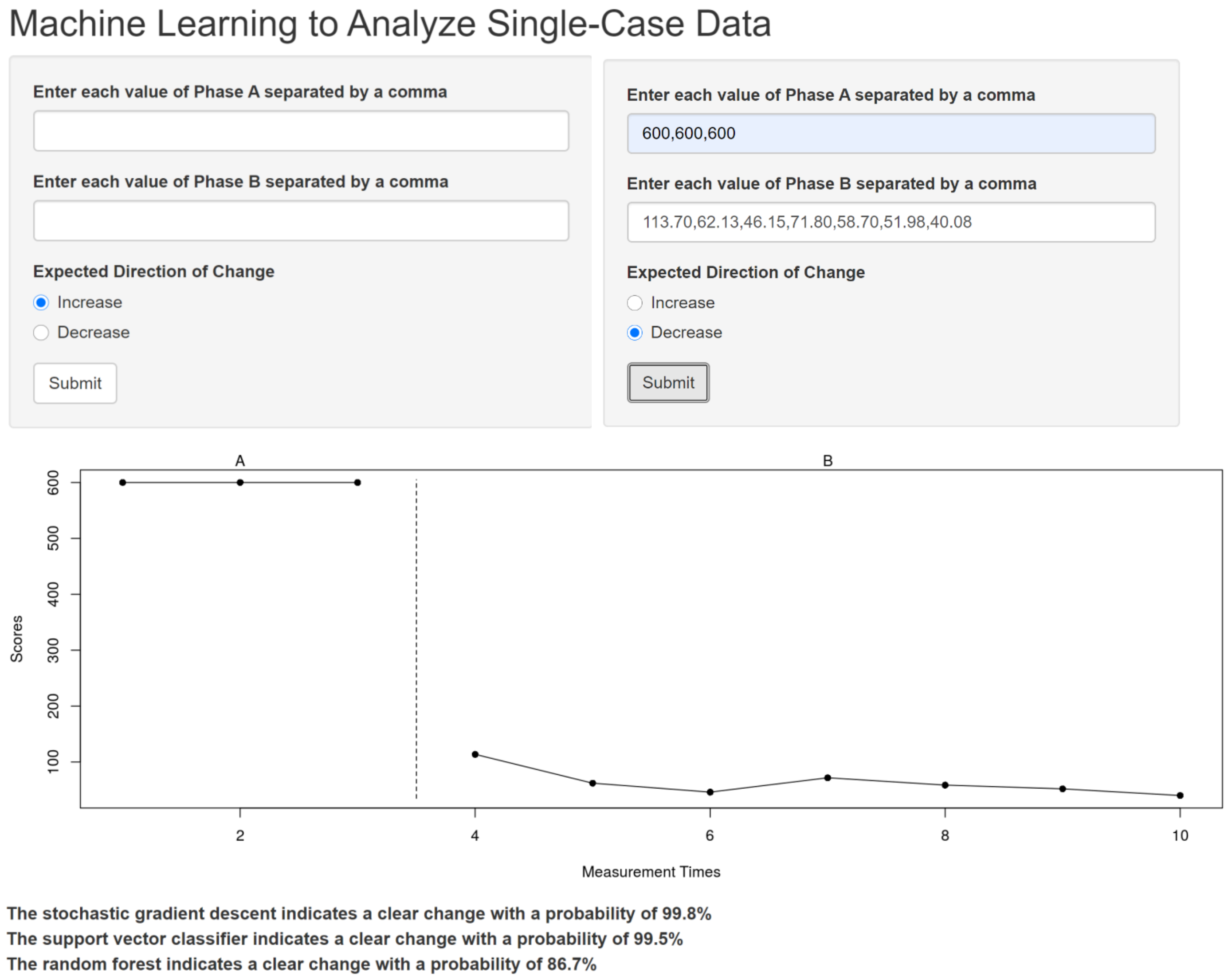
Figure 2

Refusal and Expulsion per Minute and Percentage of Session Negative Vocalisations (Top Panel) and Side Deposit and Redistribution per Minute (Bottom Panel).



Supplemental Information

Online Machine Learning Calculator Screenshot: https://labrl.shinyapps.io/MachineLearningABGraphs/



Machine Learning to Analyze Single-Case Data

Enter each value of Phase A separated by a comma

Enter each value of Phase B separated by a comma

Expected Direction of Change

● Increase

○ Decrease

Enter each value of Phase A separated by a comma: 600,600,600

Enter each value of Phase B separated by a comma: 113.70,62.13,46.15,71.80,58.70,51.98,40.08

Expected Direction of Change

○ Increase

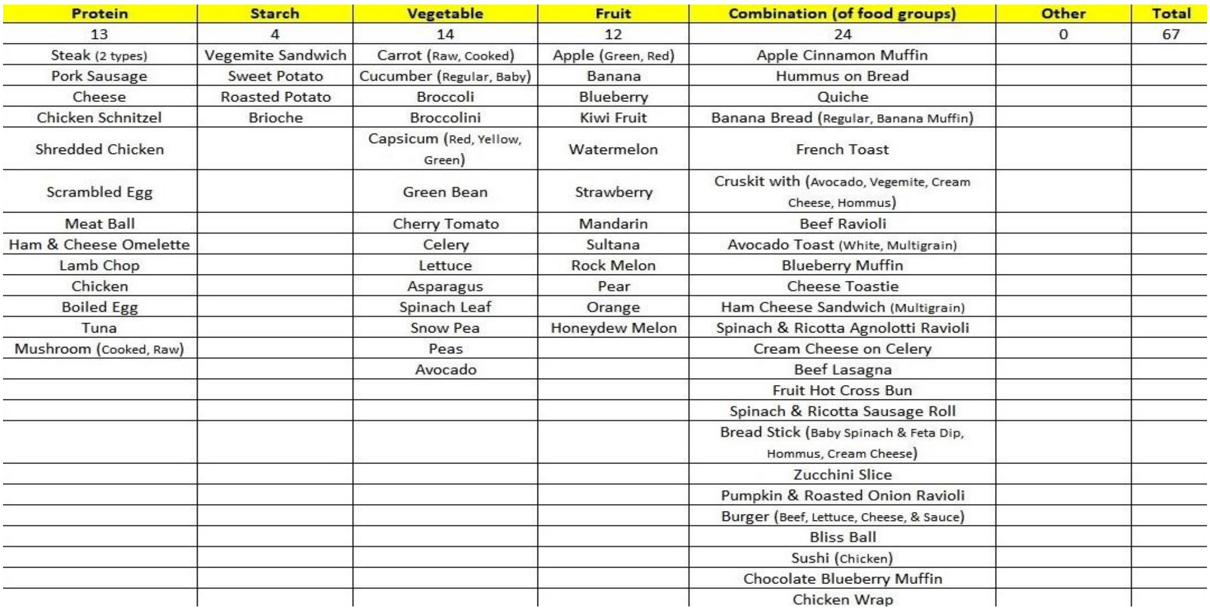
● Decrease

The stochastic gradient descent indicates a clear change with a probability of 99.8%

The support vector classifier indicates a clear change with a probability of 99.5%

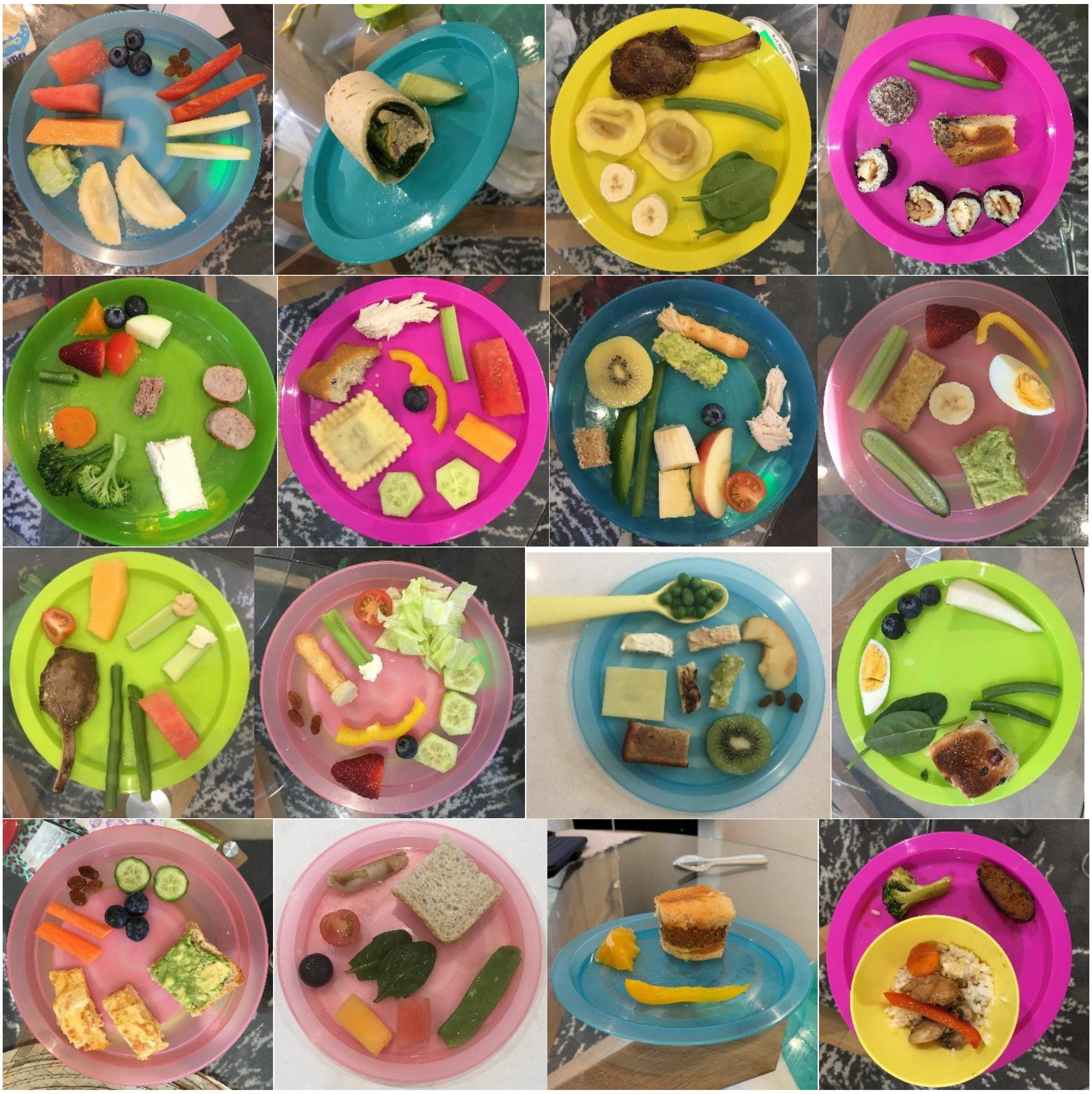
The random forest indicates a clear change with a probability of 86.7%

Julien’s Food List

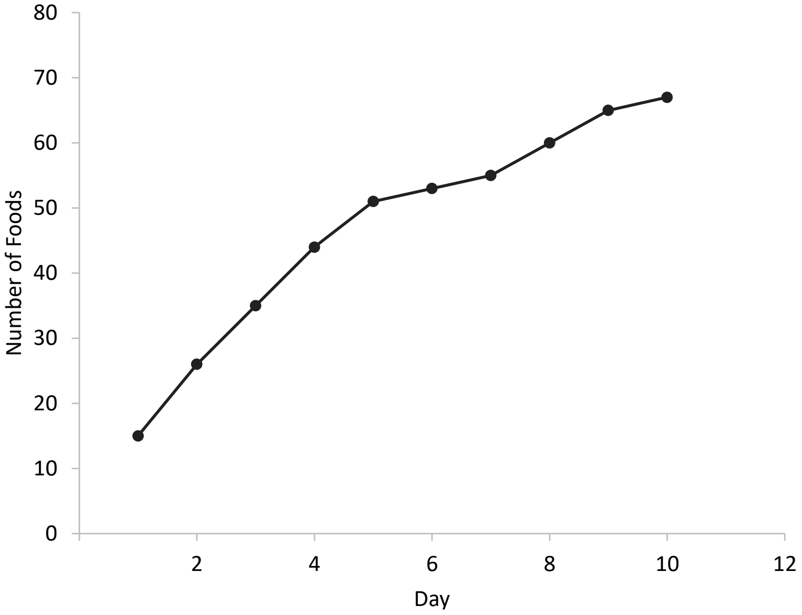


Protein	Starch	Vegetable	Fruit	Combination (of food groups)	Other	Total
13	4	14	12	24	0	67
Steak (2 types)	Vegemite Sandwich	Carrot (Raw, Cooked)	Apple (Green, Red)	Apple Cinnamon Muffin		
Pork Sausage	Sweet Potato	Cucumber (Regular, Baby)	Banana	Hummus on Bread		
Cheese	Roasted Potato	Broccoli	Blueberry	Quiche		
Chicken Schnitzel	Brioche	Broccolini	Kiwi Fruit	Banana Bread (Regular, Banana Muffin)		
Shredded Chicken		Capsicum (Red, Yellow, Green)	Watermelon	French Toast		
Scrambled Egg		Green Bean	Strawberry	Cruskit with (Avocado, Vegemite, Cream Cheese, Hommus)		
Meat Ball		Cherry Tomato	Mandarin	Beef Ravioli		
Ham & Cheese Omelette		Celery	Sultana	Avocado Toast (White, Multigrain)		
Lamb Chop		Lettuce	Rock Melon	Blueberry Muffin		
Chicken		Asparagus	Pear	Cheese Toastie		
Boiled Egg		Spinach Leaf	Orange	Ham Cheese Sandwich (Multigrain)		
Tuna		Snow Pea	Honeydew Melon	Spinach & Ricotta Agnolotti Ravioli		
Mushroom (Cooked, Raw)		Peas		Cream Cheese on Celery		
		Avocado		Beef Lasagna		
				Fruit Hot Cross Bun		
				Spinach & Ricotta Sausage Roll		
				Bread Stick (Baby Spinach & Feta Dip, Hommus, Cream Cheese)		
				Zucchini Slice		
				Pumpkin & Roasted Onion Ravioli		
				Burger (Beef, Lettuce, Cheese, & Sauce)		
				Bliss Ball		
				Sushi (Chicken)		
				Chocolate Blueberry Muffin		
				Chicken Wrap		

Julien’s Plate Pictures



Julien’s Food Variety per Day of the Programme



Julien’s Outcomes from the Intensive Paediatric Feeding Treatment Programme

Outcome Domain	Outcome
Goals Met (out of 9)	100%
Full Plate Consumption	100%
Full Plate Volume Reached	*M* = 155 grams
Texture	Full Regular Portions
Independence	Full Plate Self-feeding and Scooping with Utensils
Variety	
• Total	67
• Protein	13
• Starch	4
• Vegetable	14
• Fruit	12
• Combination	24
• Other	0
Caregiver Satisfaction	4.81 (out of 5)
Social Acceptability	4.87 (out of 5)
Percent Change and Effect Size	
• Consumption	100%
• Latency to Acceptance	98.4%
• Latency to Mouth Clean	91.5%
• Refusal	95.7%
• Expulsion	100%
• Negative Vocalisations	100%
Postdischarge Parent-fed Meals	
• Time	8 days
• Meals	*N* = 18
• Consumption	100%
• Meal Duration	*M* = 20.9 minutes (range 15 to 30)
• Meal Rating (out of 6)	5.2 (range 4 to 6)
1-year Follow-up	4 out of 5 (Much Better) All food groups; 49 out of 67 foods Zero Redistribution