Statistics in Biopharmaceutical Research

ISSN: (Print) 1946-6315 (Online) Journal homepage: https://www.tandfonline.com/loi/usbr20

Feature Selection With Weighted Importance Index in an Autism Spectrum Disorder Study

Davit Sargsyan, Shyla Jagannatha, Nikolay V. Manyakov, Andrew Skalkin, Abigail Bangerter, Seth Ness, Kathryn Durham, Dhammika Amaratunga, Javier Cabrera & Gahan Pandina

To cite this article: Davit Sargsyan, Shyla Jagannatha, Nikolay V. Manyakov, Andrew Skalkin, Abigail Bangerter, Seth Ness, Kathryn Durham, Dhammika Amaratunga, Javier Cabrera & Gahan Pandina (2019) Feature Selection With Weighted Importance Index in an Autism Spectrum Disorder Study, Statistics in Biopharmaceutical Research, 11:2, 118-125, DOI: 10.1080/19466315.2018.1537886

To link to this article: https://doi.org/10.1080/19466315.2018.1537886

Published online: 28 May 2019.

STATISTICS IN BIOPHARMACEUTICAL RESEARCH
2019, VOL. 11, NO. 2, 118–125
https://doi.org/10.1080/01621459.2018.1537886

Taylor & Francis
Taylor & Francis Group

Feature Selection With Weighted Importance Index in an Autism Spectrum Disorder Study

Davit Sargsyan[a], Shyla Jagannatha[b], Nikolay V. Manyakov[c], Andrew Skalkin[d], Abigail Bangerter[b], Seth Ness[b], Kathryn Durham[b], Dhammika Amaratunga[e], Javier Cabrera and[f], and Gahan Pandina[b]

[a]Translational Medicine and Early Development Statistics, Janssen Research & Development, LLC, Spring House, PA; [b]Janssen Research & Development, LLC, Titusville, NJ; [c]Janssen Research & Development, LLC, Beerse, Belgium; [d]Janssen Research & Development, LLC, Spring House, PA; [e]Rutgers University, New Brunswick, NJ; [f]Rutgers University, Piscataway, NJ

ABSTRACT

Elastic net regularization is a popular statistical tool for variable selection that combines lasso and ridge regression penalties. When used in combination with ensemble methods, it improves stability of the estimates and increases confidence in the results. We proposed and tested a version of this method that considers a measure of models' goodness of fit and gives estimates of importance for each feature weighted on this measure. The method was applied to an autism spectrum disorder (ASD) study to select a subset of biosensor-based features that can be used to predict clinical scores of study participants. In this study, the participants' responses to visual and audio stimuli were captured by the Janssen Autism Knowledge Engine (JAKE®) biosensors and used to construct approximately 50,000 features. We examined how well changes in these features mirrored changes in the Social Responsiveness Scale (SRS), a quantitative assessment of ASD individuals by clinicians. As a result, we isolated the top features changes which are most associated with changes in SRS, and built predictive models using these features.

ARTICLE HISTORY
Received November 2017
Revised September 2018

KEYWORDS
Data mining; Elastic net; Ensemble methods; JAKE; Lasso; Ridge regression;

1. Introduction

Autism spectrum disorder (ASD) is a variety of neurological conditions manifested through communication deficits such as inappropriate responses in conversations, misreading nonverbal messages, or difficulties in forming relationships with their peers, accompanied by repetitive or obsessive behavior and high sensitivity to changes in their environment (American Psychiatric Association 2013). Janssen Autism Knowledge Engine (JAKE®), a multimodal data capture system in individuals with ASD, was developed by Janssen Pharmaceutical Research and Development to optimize the ability to identify subpopulations for research and sensitively measure treatment outcomes (Ness et al. 2017). JAKE is a three-part investigational system consisting of: My JAKE, a web-based and mobile application tool to collect clinical data and monitor symptoms and progress of the study participants; JAKE Sense, a group of biosensors and tasks designed to assess physiological characteristics and behaviors related to core ASD symptoms; and JAKE Stream, a system designed to capture and process data from both My JAKE and JAKE Sense. JAKE Sense periodic biosensors recorded electroencephalography (EEG), electrocardiography (ECG), electrodermal activity (EDA), eye movement, and facial expressions (FACET) of the participants as they responded to visual and audio stimuli throughout a series of tests given during clinical visits; additionally, continuous measurements for daytime activity and nighttime sleep patterns were recorded daily. Combining clinical data from multiple sources, such as JAKE Sense data, often leads to situations where a large number (p) of explanatory variables, or features, exists for a relatively small number (N) of participants, that is, $p \gg N$. Here, we explored 48,036 features constructed from biosensor data collected from 129 participants with ASD over three clinical visits. We examined how changes in these features correlated with changes in clinical scores in the Social Responsiveness Scale (SRS), which measures ASD symptoms associated with general behaviors.

Social Responsiveness Scale is a quantitative measure of autistic traits in 4- to 18-year-olds that aggregates responses to a 65-item questionnaire. Clinicians evaluate each question on a scale between 0 (not true) and 3 (almost always true). The questions focus on participant's behavior in the past 6 months (Bölte et al. 2008). Total scores are then transformed into T-scores that are standardized to mean of 50 and SD of 10. Participants with T-score of 76 or higher are considered severe, and strongly associated with a clinical diagnosis of ASD; participants with scores between 60 and 75 are considered mild to moderate (Aldridge et al. 2010).

The primary goal of this analysis was to identify a subset of biosensor-based features changes that mirrored longitudinal trends in SRS (differences between first and last visits, in this case), and could be used to assess ASD participants' progress without relying on subjective measurements, that is, clinical scales.

CONTACT Davit Sargsyan dsargsy@its.jnj.com Translational Medicine and Early Development Statistics, Janssen Research & Development, LLC, 1400 McKean Road, Spring House, PA 08560.

Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/SBR.

2. Methods

The aim was to select subsets of biosensor-based features from the tasks designed to assess physiological characteristics related to core ASD symptoms that were associated with SRS scores of clinical significance (Appendix). We employed the enriched ensembles (EnrE) method (Amaratunga et al. 2012), combined with elastic-net-penalized generalized linear regression models (Friedman et al. 2010). We subsequently improved this combined method by weighting on goodness of fit of each model.

2.1. Elastic Net

Lasso minimizes the residual sum of squares (where residuals are the differences between the observed values and values predicted by a model) while constraining the sum of the absolute values of regression coefficients to less than a constant (Tibshirani (1996)). This allows shrinking of some coefficients to zero, that is, lasso effectively performs feature selection by removing features that are not strongly associated with the response by estimating their contribution to the model as zero. Regarding a group of highly correlated features, lasso normally picks up one while shrinking the rest of the coefficients for the group to zero. This creates a possibility of multiple or unstable solutions.

In contrast, ridge regression pulls coefficients of a group of highly correlated features closer together by constraining the sum of squares of these coefficients to less than a constant while minimizing the sum of squares of residuals. As a result of this procedure, the obtained estimates of regression coefficients are corrected for nonorthogonality of the features' space. Although solutions here are stable, ridge regression does not reduce the number of features.

Elastic net is a compromise between the lasso (L1) and ridge regression (L2) penalties. This mixed approach is especially useful for datasets with highly correlated features. To illustrate the method, we adopt the following notations:

Given a vector of response variable

$$Y = [y_1, y_2, \ldots, y_N]^T$$

and matrix of features

$$X = \begin{bmatrix} x_{11}x_{12} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{N1}x_{N2} & \cdots & x_{Np} \end{bmatrix}$$

we want to estimate a vector of coefficients

$$\beta = [\beta_1, \beta_2, \ldots, \beta_p]^T, \quad \text{with} \quad p \gg N.$$

By balancing the two aforementioned regularization techniques, it is then possible to minimize the following expression:

$$\frac{1}{2N}\sum_{i=1}^{N}(y_i - \beta_0 - x_i\beta)^2 + \lambda h(\alpha, \beta)$$

Where $x_i = [x_{i1}, x_{i2}, \ldots, x_{ip}]$, $h(\alpha, \beta)$ is a function of L1 and L2 penalties, and α is the balance between them:

$$h(\alpha, \beta) = \sum_{j=1}^{p}[\frac{1}{2}(1-\alpha)\beta_j^2 + \alpha|\beta_j|]$$

When α is zero, a ridge regression is performed; on the other extreme, when α is one, the method is equivalent to lasso. By adjusting α in the range between 0 and 1, a solution can be found such that only a subset of features is kept (i.e., has nonzero coefficients) while the procedure simultaneously corrects for features' correlations.

One must note that opinions are divided regarding the optimization, or "tuning" the parameter α. In our opinion, this should never be done; rather, the balance between the two penalties should be decided by trying a handful of values of α based on the desirable magnitude of shrinkage as it is highly dependent on the correlation structure of the data

2.2. Enriched Ensembles

Enriched ensembles method is largely an improvement on bagging-only ensemble (BagE) and simple ensemble (SimE) methods. The data is initially divided into a training and a testing set (usually at the 80%/20% rate by participants). Low signal in the data is amplified by weighted sampling of $p^* = \sqrt{p}$ of number of features for $N^* = 0.63N$ of the participants from the training set (the coefficient 0.63 here mimics bootstrap sampling effect). The sampling weights' vector is a function of p-values calculated for each feature in relationship to the response, individually. An argument was made for using false discovery rate (FDR)-adjusted p-values instead of row p-values to reduce overfitting (Amaratunga et al. 2012). The sampling is done repeatedly several hundred to several thousand times, creating a new subset of subject and features every time. A model is fitted to each sample, and the results of each fit are recorded. The results are aggregated at the end and the final model applied to the testing set.

Generally, when p is much larger than N (as a rule of thumb, when the ratio $N/log(p)$ is less than 20), the number of linear combinations becomes so large that it overwhelms the model resulting in poor performance. By repeatedly sampling the data and using small subsets to fit the model, we reduce the chance of overfitting. A similar philosophy employed in gene sequencing, for example, where instead of trying to process large chunks of material at once, long sequences are broken into much smaller pieces, processed, and reassembled at the end.

In our study, we applied this technique to compute the importance of each feature in relationship to the response. Feature importance index (FII) was defined as the ratio of number of times a feature was selected by the model over the number of times that feature was sampled:

$$\mathrm{FII}_j = \frac{\sum_{k=1}^{K} C_{kj}}{\sum_{k=1}^{K} S_{kj}}$$

where S_k and $C_m$are respective indicators of jth feature being present in the kth sample (*s*ampled) and subsequently being selected by the model (*c*hosen, i.e., having a nonzero coefficient as estimated by elastic net), with values of 0 or 1. K is the total number of samples taken.

One immediate consequence of this calculation is that the contribution of each sample is equivalent; hence, simply being selected by the model is the only metric that matters for FII. In reality, samples can vary vastly which will be reflected in the way data fits to the models.

Table 1. Number of features and participants analyzed in each experiment.

Experiment	Number of features	Missing >30%	Number of features analyzed	Number of participants analyzed
Activity monitoring	287	161	126	120
Biomotion	11,853	11,747	106	120
Social information processing	804	708	96	123
ER	2247	1874	373	121
Facial expression production	342		342	99
Funny videos	684	627	57	103
HourGlass	84		84	120
Dynamic videos	14,624	14,326	298	119
Emotional face	12,536	12,440	96	121
Visual search	607	417	190	114
TOTAL	48,036			

NOTE: ERP, event related potentials; VET, visual exploration task.

2.3. Weighted Index of Feature Importance

To correct for the fact that not all samples will be a good representation of the data, we proposed and implemented a weighted version of importance index. For each sample, a goodness of fit measure was collected along with the list of features sampled and selected (i.e., with nonzero regression coefficients). The weighted index of feature importance (WIFI) was then calculated for each feature ($j = 1, \ldots, p$) as following:

$$\mathrm{WIFI}_j = \frac{\sum_{k=1}^{K} w_k C_{kj}}{\sum_{k=1}^{K} w_k S_{kj}}$$

where w_k is a function of goodness of fit in the model applied to the k^{th} sample.

2.4. Elastic net within EnrE framework

Applying the methods described above to our dataset, we broke the analysis into several steps as described below.

Step 1. Data pre-processing

Raw data containing a total of 54,287 features constructed from biosensor records of 184 participants was processed by removing all typically developing participants (who had only a baseline visit), ASD participants who did not complete all three visits, and features that were constructed using data from only a part of an experiment. The remaining data contained 48,036 features for 129 ASD participants.

Next, we separated the data by the experiments (see Table 1 for details). Within each experiment, we computed first-to-last visit differences for each feature and each subject. Features with near-zero variance, defined by default as features with less than 10% unique values and more than 95:5 ratio of frequencies of the most common value and the second most common value (Kuhn 2008), were removed. Additionally, features with more than 30% missing data, and participants with more than 80% missing data were excluded. The data was then centered and scaled, and the remaining missing values were imputed using k-nearest neighbors (KNN) algorithm (Altman 1992).

To identify and address potential outliers, we used principal component analysis (PCA) followed by the modified z-score (MedMAD) test (Iglewicz and Hoaglin 1993). Using the first two principal components, MedMAD test flagged participants with z-scores beyond the 3.5 threshold recommended by Iglewicz and Hoaglin and excluded them from the analysis (see Figure 1 for graphical representation of data before and after outlier removal).

Step 2. Separate training and testing sets

The data was divided into a training set and a testing set, with roughly 80% (N^*) of participants selected into the training set at random and without replacement.

Step 3. Elastic net

(a) Linear regression models were fitted for the response (differences in SRS between first and last visits) versus each feature in the training set. The weight for each feature was then calculated as a function of these univariate p-values as following:

$$u_j = \ln(q_j) / \sum_{j=1}^{p} \ln(q_j),$$

where q_j is the p-value for the jth feature adjusted for multiplicity following the FDR method (Benjamini and Hochberg 1995).

(b) A random sample of $0.63N^*$ of the training set's participants and $\sqrt{p}$ of number of features was taken without replacement.

(c) Generalized linear model (GLM) with elastic net penalties was fitted to the subset using 10-fold cross-validation (Picard and Cook 1984). Multiple values of the mixing parameter α were tested before deciding on 0.5 (i.e., a 50/50 mix of lasso and ridge regression) rather than formally searching for an optimal balance. At this stage, an optimal value of the parameter λ was determined as the smallest λ that minimized the cross-validation error.

(d) Given the optimal value of λ, we extracted the list of features present in the sample and the list of features with nonzero coefficients as estimated by the model, as well as the goodness of fit estimates. We used the deviance ratio: the fraction of deviance explained by the saturated model compared to the null (intercept-only) model, as an assessment of goodness of fit:

$$\mathrm{dev} = 2\left[\mathrm{loglike}\,(M_s) - \mathrm{loglike}\,(M_0)\right],$$
$$\text{hence dev.ratio} = \frac{1-\mathrm{dev}}{\mathrm{dev}_{\mathrm{null}}}$$

(e) We repeated Steps 3–5, 10,000 times (given the high number of features, we set the number of samples to 10,000 to allow more features a chance to appear in the models).

(f) The output from all iterations was aggregated using the WIFI index described above. The features were then sorted by importance corresponding to their WIFI values.

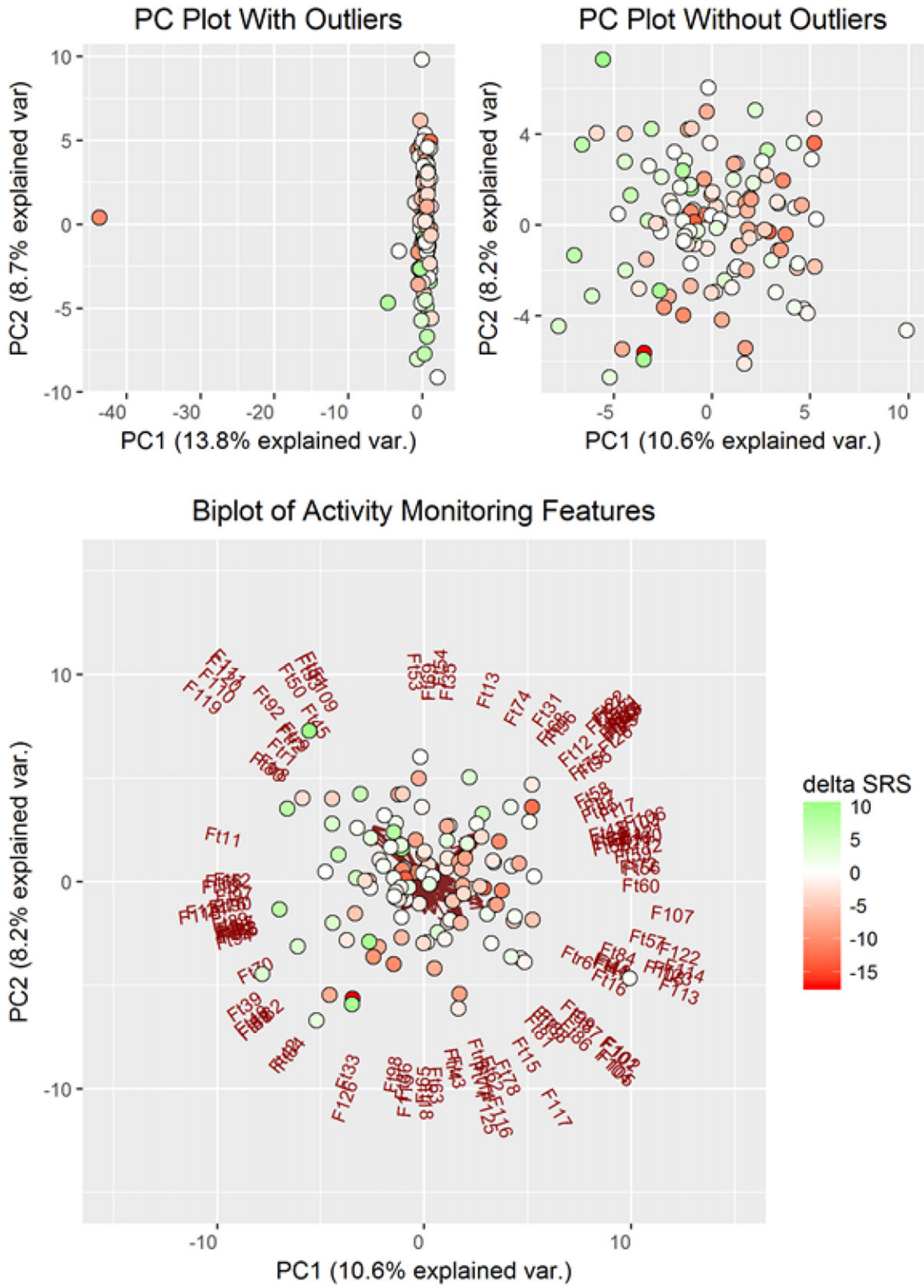

Figure 1. First two principal components of the differences in features between the last timepoint and baseline for the activity monitoring experiments: (upper left) before; (upper right) after removal of outliers; (lower) biplot of the first two principal components, outliers removed. *Ft1* to *F126* is shortened for Feature 1–126.

Sequential addition of features in decreasing order of the estimated importance was used to construct regression models on the training set. The models were compared two at time, and a more complex model was selected if it significantly improved the fit. The process continued until no further significant improvement was observed and the simplest of the two models was selected. Based on the model comparison results, a model with the top three features was selected. The final model was then fit to the testing set.

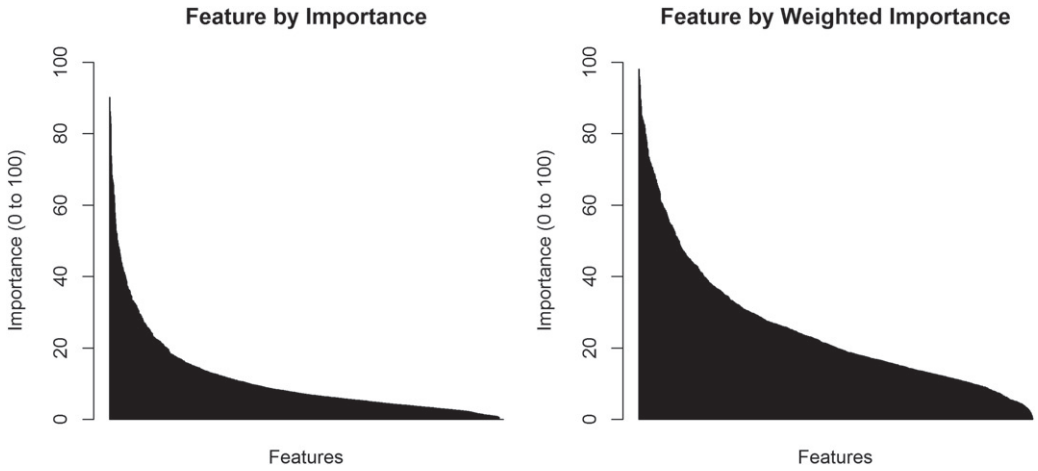

Figure 2. Importance indices of 1768 biosensor-based features (10 experiments combined).

Table 2. Feature importance.

Feature	Univariable coefficient	Unweighted importance	Weighted importance
bm | average AU23 Evidence | baseline not corrected	−2.08	0.88	0.98
Activity monitoring | all videos together | % furniture (to total stimulation time) [%]	−2.36	0.90	0.97
Facial expression production | sad | area under fear evidence curve	1.99	0.86	0.96
bm | average AU23 evidence | baseline corrected	−2.29	0.83	0.95
Activity monitoring | all videos together | % furniture (to total valid time) [%]	−2.41	0.85	0.95
Facial expression production | happy | sum variance neutral evidence	−1.94	0.84	0.94
Social versus nonsocial face | average AU7 evidence | baseline corrected	1.67	0.67	0.91
Social versus nonsocial face_obj | average negative evidence | baseline corrected	−1.83	0.74	0.90
Facial expression production | surprised | average surprise evidence	−1.78	0.72	0.89
Social versus nonsocial face | average negative evidence | baseline corrected	−2.02	0.67	0.88

NOTE: bm, biomotion.

3. Results

Elastic net within EnrE was applied to the recombined dataset with 1768 features across 10 experiments and 83 participants using the algorithm described in the methods section. 64 participants were selected at random into the training set, with 43 participants sampled without replacement at each iteration. The weighted and unweighted importance indices were calculated for each feature. The indices in decreasing order are presented in Figure 2.

Generally, weighed importance indices were higher than corresponding unweighted indices. Table 2 shows the top 10 features by the weighted importance index.

Using the training dataset, features were added to regression models sequentially, in order of decreasing weighted importance, and the models compared using ANOVA. Fit improved significantly for the first three features, increasing R^2 from 0.137 in a model with a single feature (biomotion, average AU23 evidence without baseline correction) as a predictor of change in SRS, to 0.361 in a model with top three features. This final model was further validated with the testing set using data from the remaining 19 participants. Changes in SRS predicted by the final model using the testing set were compared against the observed SRS changes in the testing set. The correlation between the predicted and the observed SR changes was low (0.12) as shown in Figure 3. These results were not unexpected, the methodology was looking for meaningful changes over time. However, the study was observational in nature, that is, there was no study-specific intervention. The changes in SRS measure, therefore, carried large amount of noise (Figure 4). The method, however, was able to detect the difference in feature values between the last time point and baseline.

We also compared weighted and unweighted methods using simulated data. The simulations show that on average, WIFI is more conservative than the unweighted method. Figure 5 shows the results of simulation with varying signal-to-noise ratios. In this example, 1% of the features were correlated with the response.

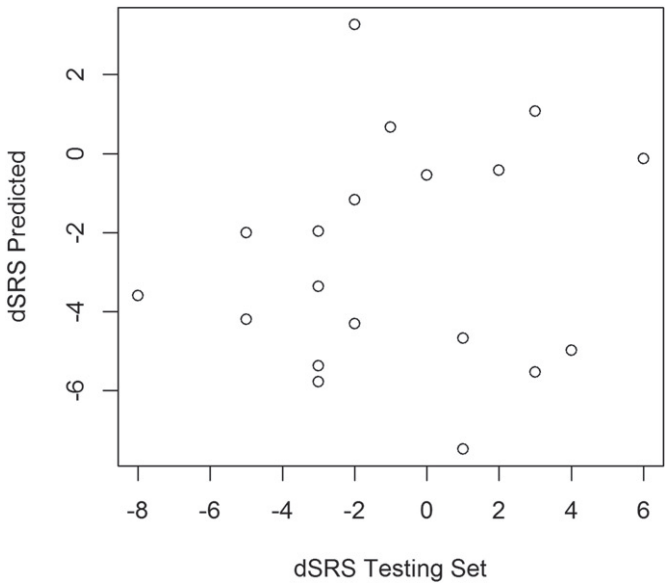

Figure 3. Change in SRS, observed versus predicted using testing set.

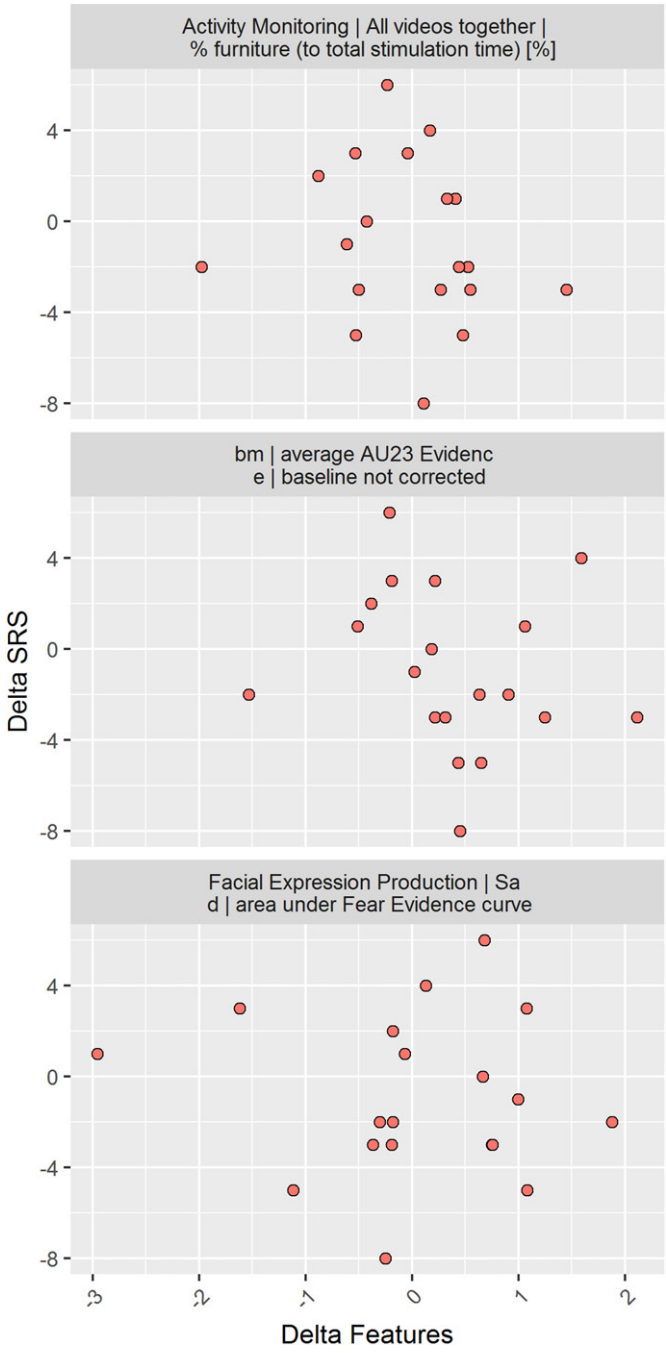

Figure 4. Changes in top three features versus changes in SRS, first-to-last visit.

$x_{ip} = \beta * y_{ip} + \varepsilon_i$, where y_{ip} and ε_i are sampled from normal distribution with mean 0 and SD 1.

The features were generally ranked lower by WIFI compared to univariable p-value and unweighted importance ranking.

Further validation of the method using data from more studies and simulations, possibly with larger sample size, are necessary to identify scenarios where WIFI has an advantage over unweighted version of the feature selection method.

4. Discussion

The use of objective sensor-based data obtained via JAKE in individuals with ASD has the potential to transform the understanding of the pathophysiology of ASD, enhance assessments, aid in subtype identification, individualize treatments, and track developmental outcomes. In this study, we were able to identify a subset of biosensor-based features with changes that mirrored trends in clinical scores. This identification was accomplished by applying data mining techniques that combined penalized regression with EnrE sampling method, and goodness of fit measures of to give higher weights to features selected from models that better described the observations.

As an emerging field, a limited number of reports are available on the use of biosensors to help identify diagnostic features in individuals with ASD. Recent studies have demonstrated the usefulness of eye-tracking measures for classification and aspects of social interaction performance (Hanley et al. 2015;

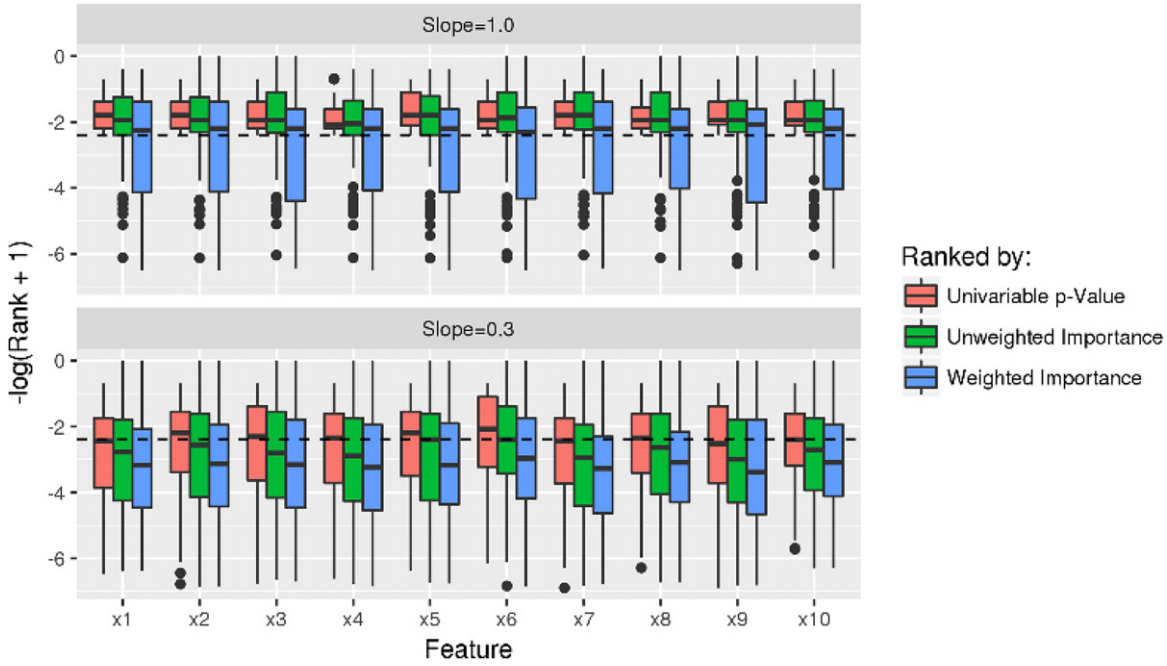

Figure 5. Ranking of features correlated with the response. The simulated data contained 1000 features, 10 of which were correlated with the response. The results are from 100 simulated datasets.

Frazier et al. 2016). Use of an eye tracking-based autism risk index demonstrated that this index had substantial diagnostic accuracy, outperformed the SRS, and showed strong relationships with a gold-standard measure of autism symptom severity (Frazier et al. 2016). In a study using a different objective measurement, it was found that pupil dilation metrics correlated with individual differences measured by the SRS (DiCriscio and Troiani 2017).

Extracting meaningful information from such complex data requires understanding of the data structure and thorough examination of relationships that exist between individual components. Statistical methods used in this analysis generally perform better when the data come from a single source, for example, gene expression data. We grouped features by experiments to make data more homogeneous within each group but disregarded biosensors that were involved in capturing the data. However, since data quality varied for different biosensors, with, for example, eye-tracker, sleep monitor, and ECG capturing higher quality signals compared to EEG (Ness et al. 2017), more in-depth analysis might be necessary. One possibility is to add a grouping variable for the biosensor and either incorporate it into the models or perform a subgroup analysis for each biosensor separately. Another potential direction is to use models that are robust for outliers. Although we excluded participants that were revealed by PCA as influential points, there are still questions about the distributions and outliers on the level of individual features. In such cases, downweighing of outliers might be used as an alternative to exclusion.

Acknowledgments

The authors thank Stacey E. Shehin, PhD (PRA Health Sciences) and Ellen Baum, PhD (Janssen Global Medical Affairs) for writing and editorial assistance, respectively. This study was funded by Janssen Research & Development, LLC. Davit Sargsyan, Shyla Jagannatha, Nikolay V. Manyakov, Andrew Skalkin, Abigail Bangerter, Seth Ness, and Gahan Pandina are employees of Janssen Research & Development, LLC, and may hold stock options or shares in the company. Kathryn Durham was a contractor for Janssen Research & Development at the time of the study. Dhammika Amaratunga and Javier Cabrera report no conflicts of interest. All authors meet ICMJE criteria, had full access to the study data, and take responsibility for integrity of the data. All authors have approved the final manuscript. Portions of this study have been presented at the ASA Biopharmaceutical Section Nonclinical Biostatistics Conference, June 12th–14th, 2017, Rutgers University, Piscataway, NJ.

Appendix: The tasks and stimuli used with the biosensors

Task	Description
Resting state	Video of sand falling through an hourglass for 1 min
Resting state—eyes closed	Subject asked to close their eyes for 45 sec
Event related potentials (ERP)	Measure of brain activity in response to static facial stimuli with averted or direct gaze
Social information processing	Video of male or female actor presented in random order Actor engages subject in direct speech (dyadic bid) and joint attention, toward or away from a moving toy
Dynamic videos	Measure of brain activity in response to dynamic videos of children's faces (social) or toys (nonsocial)
Visual search	Free viewing of Arrays of 24 images (including social images, HAI and low autism interest [LAI] objects
Biological motion	Two side by side videos, in random left-right order. Each video contains dynamic point-light displays. One video is derived from human actor's performance; the other video is a computer-generated animation of moving dots.
NimStim emotional faces	Measure of brain activity in response to happy, angry, fear and neutral faces
Activity monitoring	Video recording of multiple human actors performing a social activity, with visually salient distracters in the background. Actors focus on each other or on the activity only in two conditions
Funny videos	Funny videos, or videos designed to elicit an emotional response of surprise or joy were shown
Emotional faces expression	Subjects will also be asked to make faces to reflect basic emotions: Happy, Sad, Surprise, Scared, Angry, Yucky (disgust).
Auditory stimuli	Three sets presented auditory stimuli (toilet flush, a ticking clock, or an 880 hz tone). Sounds were presented for 3 sec duration, with ISI of 8–12 s. Screen displayed bubbles screen saver and a progress bar indicating time until the presentation was complete.

References

Aldridge, F., Schmidhofer, K., Gibbs, V., and Williams, M. (2010), "Evaluating the Usefulness of the Social Responsiveness Scale (SRS)," Australian Autism Alliance, available at *https://www.autismspectrum.org.au/sites/default/files/PDFuploads/Research%20Insights%20News%20Iss%203-LR.pdf*. [118]

Altman, N. S. (1992), "An Introduction to Kernel and Nearest-Neighbor Nonparametric Regression," *The American Statistician*, 46, 175–185. [120]

Amaratunga, D., Cabrera, J., Cherckas, Y., and Lee, Y.-S. (2012), "Ensemble Classifiers," *Institute of Mathematical Statistics Collections*, 8, 235–246. [119]

American Psychiatric Association (2013), *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.), Arlington, TX: American Psychiatric Publishing. [118]

Benjamini, Y., and Hochberg, Y. (1995), "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing," *Journal of the Royal Statistical Society* Series B, 57, 289–300. [120]

Bölte, S., Poustka, F., and Constantino, J. N. (2008), "Assessing Autistic Traits: Cross-Cultural Validation of the Social Responsiveness Scale (SRS)," *Autism Research*, 1, 354–363. [118]

DiCriscio, A. S., and Troiani, V. (2017), "Pupil Adaptation Corresponds to Quantitative Measures of Autism Traits in Children," *Scientific Reports*, 7, 6476. [124]

Frazier, T. W., Klingemier, E. W., Beukemann, M., Speer, L., Markowitz, L., Parikh, S., Wexberg, S., Giuliano, K., Schulte, E., Delahunty, C., Ahuja, V., Eng, C., Manos, M. J., Hardan, A. Y., Youngstrom, E. A., and Strauss, M. S. (2016), "Development of an Objective Autism Risk Index Using Remote Eye Tracking," *Journal of the American Academy of Child & Adolescent Psychiatry*, 55, 301–309. [124]

Friedman, J., Hastie, T., and Tibshirani, R. (2010), "Regularization Paths for Generalized Linear Models via Coordinate Descent," *Journal of Statistical Software*, 33, 31–22. [119]

Hanley, M., Riby, D. M., Carty, C., Melaugh McAteer, A., Kennedy, A., and McPhillips, M. (2015), "The Use of Eye-Tracking to Explore Social Difficulties in Cognitively Able Students with Autism Spectrum Disorder: A Pilot Investigation," *Autism*, 19, 868–873. [123]

Iglewicz, B., and Hoaglin, D. (1993), "The ASQC Basic References in Quality Control: Statistical Technique" in *How to Detect and Handle Outliers* (Vol. 16), Milwaukee, WI: ASQC Quality Press. [120]

Kuhn, M. (2008), "Building Predictive Models in R Using the Caret Package," *Journal of Statistical Software*, 28, 1–26. [120]

Ness, S. L., Manyakov, N. V., Abigail, B., Lewin, D., Jagannatha, S., Boice, M., Skalkin, A., and Pandina, G. (2017), "JAKE® Multimodal Data Capture System: Insights from an Observational Study of Autism Spectrum Disorder," *Frontiers in Neuroscience*, 11, 517. [118,124]

Picard, R., and Cook, D. (1984), "Cross-Validation of Regression Models," *Journal of the American Statistical Association*, 79, 575–583. [120]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [119]