




# Hold-out validation for the assessment of stability and reliability of multivariable regression demonstrated with magnetic resonance imaging of patients with schizophrenia

Jacob Levman<sup>1,2</sup>  | Maxwell Jennings<sup>1,3</sup>  | Priya Kabaria<sup>4</sup> | Ethan Rouse<sup>1</sup> | Masahito Nangaku<sup>4</sup> | Derek Berger<sup>1</sup>  | Iker Gondra<sup>1</sup> | Emi Takahashi<sup>4</sup> | Pascal Tyrrell<sup>5,6,7</sup>

<sup>1</sup>Department of Computer Science, St. Francis Xavier University, Antigonish, Nova Scotia, Canada

<sup>2</sup>Canada Research Chair in Bioinformatics, St. Francis Xavier University, Antigonish, Nova Scotia, Canada

<sup>3</sup>Department of Mathematics and Statistics, St. Francis Xavier University, Antigonish, Nova Scotia, Canada

<sup>4</sup>Department of Medicine, Boston Children's Hospital, Harvard Medical School, Boston, Massachusetts, USA

<sup>5</sup>Department of Medical Imaging, University of Toronto, Toronto, Ontario, Canada

<sup>6</sup>Department of Statistical Sciences, University of Toronto, Toronto, Ontario, Canada

<sup>7</sup>Institute of Medical Science, University of Toronto, Toronto, Ontario, Canada

## Correspondence

Jacob Levman, PhD, Canada Research Chair in Bioinformatics, St. Francis Xavier University, Antigonish, NS B2G 2W5, Canada.  
Email: jlevman@stfx.ca

## Funding information

Nova Scotia Research and Innovation Trust, Grant/Award Number: R0176004; St. Francis Xavier University, Grant/Award Number: R0168020; Natural Sciences and Engineering Research Council of Canada;

## Abstract

Neuroscience studies are very often tasked with identifying measurable differences between two groups of subjects, typically one group with a pathological condition and one group representing control subjects. It is often expected that the measurements acquired for comparing groups are also affected by a variety of additional patient characteristics such as sex, age, and comorbidities. Multivariable regression (MVR) is a statistical analysis technique commonly employed in neuroscience studies to “control for” or “adjust for” secondary effects (such as sex, age, and comorbidities) in order to ensure that the main study findings are focused on actual differences between the groups of interest associated with the condition under investigation. It is common practice in the neuroscience literature to utilize MVR to control for secondary effects; however, at present, it is not typically possible to assess whether the MVR adjustments correct for more error than they introduce. In common neuroscience practice, MVR models are not validated and no attempt to characterize deficiencies in the MVR model is made. In this article, we demonstrate how standard hold-out validation techniques (commonly used in machine learning analyses) that involve repeatedly randomly dividing datasets into training and testing samples can be adapted to the assessment of stability and reliability of MVR models with a publicly available neurological magnetic resonance imaging (MRI) dataset of patients with schizophrenia. Results demonstrate that MVR can introduce measurement error up to 30.06% and, on average across all considered measurements, introduce 9.84% error on this dataset. When hold-out validated MVR does not agree with the results of the standard use of MVR, the use of MVR in the given application is unstable. Thus, this paper helps evaluate the extent to which the simplistic use of MVR introduces study error in neuroscientific analyses with an analysis of patients with schizophrenia.

**Abbreviations:** BCH, Boston Children's Hospital; MCIC, Mind Clinical Imaging Consortium; MRI, magnetic resonance imaging; MVR, multivariable regression.

Jacob Levman and Maxwell Jennings are co-primary authors.

Canada Foundation for Innovation;  
Canada Research Chairs, Grant/Award  
Number: 231266; National Institutes of  
Health, Grant/Award Numbers:  
R01NS109475, R03NS091587,  
R21MH118739

**KEYWORDS**

hold-out validation, magnetic resonance imaging, multivariable regression

## 1 | INTRODUCTION

Neuroscience studies often involve a comparison between two groups of patients, typically one with a specific pathological condition and one control/healthy group, to determine if there is a measurable difference between the distributions of biomarker measurements acquired. It is common for many of these acquired biomarker measurements to be influenced by additional patient characteristics such as age and sex. It is widely considered undesirable to have an assessment of a given measurement's ability to characterize a pathological condition of interest influenced by secondary factors such as age and sex effects. Multivariable regression (MVR) is a statistical analysis technique commonly employed in neuroscience studies to control for (or adjust for) the secondary effects associated with sex, age, comorbidities, and so forth. The regression procedure produces a modified measurement for comparing groups that attempts to ignore secondary effects when assessing a measurement's ability to characterize a pathological condition. Although MVR is commonly used, it is not well understood whether the adjustments made by MVR correct for more error than they introduce; furthermore, no effort is made in typical neuroscience studies to characterize the amount of error introduced by the use of MVR, nor to assess MVR's stability in the given application. Additionally, it has been previously demonstrated that random guessing can outperform the least squares estimator (Davis-Stober & Dana, 2014), which is commonly used in MVR techniques, implying that MVR may be considerably less reliable in its application in standard neuroscientific analyses than many neuroscience researchers appreciate. The work presented in this manuscript attempts to assess the reliability and stability associated with the use of MVR using common machine learning techniques for model validation and, to the best of our knowledge, is the first such study to focus on patients with schizophrenia.

From a functional technical perspective, the algorithms used to perform MVR are very similar to supervised machine learning algorithms running in regression mode: both algorithms require an array of sample measurements, and both techniques similarly compute regression statistics. Unlike machine learning,

multivariable linear regression is used to model a linear relationship between a set of independent variables and the dependent variable (Alexopoulos, 2010). Very recent studies have proposed the application of concepts from prediction and machine learning to neuroscience and psychology (Bzdok, 2017; Bzdok & Ioannidis, 2019; Bzdok & Meyer-Lindenberg, 2018; Yarkoni & Westfall, 2017); however, the vast majority of the neuroscience literature does not address the issue of MVR stability. In this paper, we investigate the potential application of machine learning validation techniques to MVR model evaluation as a means to assess the reliability/stability of regression performed by MVR in the context of neuroscientific analyses. We hypothesize that MVR models are capable of overfitting to the schizophrenia data on which the model is constructed and thus resultant overfitted models will be unreliable and not robust when regressing samples not relied upon in model construction. In this paper, we investigate the potential for using the hold-out method (randomly dividing our data into training and testing sets) for MVR model validation and demonstrate the importance of doing so using real-world data commonly studied using MVR techniques. We selected a publicly available dataset of magnetic resonance imaging (MRI) examinations containing images acquired from patients with schizophrenia as well as from healthy controls. The techniques investigated in this study can be implemented as a validation tool allowing researchers to estimate the reliability and stability associated with the traditional use of MVR in a given application.

Previous research in the application of machine learning to schizophrenia has focused on diagnostic applications responsible for discriminating patients with schizophrenia from those with bipolar disorder (Schnack et al., 2014) and healthy controls (Iwabuchi et al., 2013; Nieuwenhuis et al., 2012; Schnack et al., 2014). Research has also focused on using machine learning to classify childhood-onset schizophrenia (Greenstein et al., 2012). Additional research on patients with schizophrenia has focused on the differential diagnosis between those with and without auditory hallucinations using resting state functional MRI (Chyzyk et al., 2015) as well as classifying patients into cognitive subtypes (Gould et al., 2014). Combining MRI and genetic data to improve

classification of schizophrenia has also been investigated (Yang et al., 2010). Machine learning applied to MRI exams of patients with schizophrenia has also been the subject of a review (Veronese et al., 2013) and a meta-analysis (Kambeitz et al., 2015). In this article, we hypothesize that MVR models are capable of overfitting to the provided data on which the model is constructed and thus resultant overfitted models will be unreliable and not robust when regressing samples not relied upon in model construction in a neuroscientific analysis of patients with schizophrenia.

## 2 | METHODS

### 2.1 | Participants, data acquisition, and preprocessing

Following approval for retrospective analyses by Boston Children's Hospital's (BCH's) Institutional Review Board, the Mind Clinical Imaging Consortium (MCIC) medical imaging electronic database was accessed (Gollub et al., 2013; <https://coins.trendscenter.org/>), and all examinations with clinical data available from the public MCIC dataset were downloaded for further analysis. Written informed consent was obtained from all study participants (Gollub et al., 2013). Imaging was performed with MRI scanners across multiple imaging centers with detailed MRI protocol descriptions available in the literature (Gollub et al., 2013). Imaging was divided across multiple imaging centers employing 1.5T Siemens Sonata, 3T Siemens Trio, and 1.5T GE Signa MRI scanners with eight channel or circular polarized head coils (Gollub et al., 2013).

Each T1 structural examination was processed with FreeSurfer v6.0 (Desikan et al., 2006; Fischl, 2012; Fischl et al., 2004; Fischl & Dale, 2000; Ségonne et al., 2007) (<http://surfer.nmr.mgh.harvard.edu/>), using the recon-all command that aligns the input examination to all available brain atlases. Each FreeSurfer output T1 structural MRI examination was displayed with label map overlays and visually inspected for quality of regional segmentations. If FreeSurfer results were observed to substantially fail, they were excluded from this analysis (i.e., FreeSurfer regions of interest that do not align to the MRI). This resulted in a collection of 174 MRI examinations that passed quality control (from 213 examinations publicly accessed), including 99 patients with schizophrenia and 75 examinations of healthy control participants. All available FreeSurfer measurement types were extracted from each MRI examination (cortical thicknesses, volumes, surface areas, surface curvatures, signal intensity measurements, etc.).

### 2.2 | Statistical analysis

This study included the acquisition of 4784 regionally distributed measurements per imaging examination, as extracted by FreeSurfer's recon-all command, which processes the participant's examination with all available brain atlases (Fischl, 2012). Measurements for which invalid numbers were extracted from FreeSurfer were excluded from the analysis (i.e., not a number of values or lack of a computed measurement in the corresponding .stats file), resulting in 4482 measurements included for further analysis (this includes total brain volume, regional brain volumes, regional surface areas, and regional cortical thicknesses).

An MVR model (using MATLAB's mvregress function) (Matlab R2018a, Natick, MA, USA) was constructed for each measurement under consideration to control for (or adjust for) the effects of sex and age, projecting each measurement from each sample into the regressed data space. This represents common practice in neuroscience studies, creating adjusted versions of each sample's measurements in order for the direct comparison between groups (in this case schizophrenia vs. healthy) to attempt to avoid the effects of age and sex (which are known to affect the presentation of the brain as observed on MRI examinations) present in the current measurement under consideration.

Hold-out (machine learning style) validation was then employed to repeatedly construct MVR models with held-out training data based on random sampling. This involves a standard machine learning bootstrap validation loop in which the samples available are randomly sampled such that 50% of the randomly selected samples are used to construct the MVR model (training) and the remaining 50% of samples that were not used in constructing the MVR model are regressed into the resultant data space (testing). In order to confirm that our primary findings are not dependent on the selection of validation technique, we repeated the analysis with  $K$ -fold cross validation ( $K = 10$ ), resulting in 90% of the samples available contributing to creating our MVR models. These validation procedures (both 50/50 and  $K$ -fold) were repeated 250 times. This results in a distribution of regressed values for each measurement from each patient. We then compare the regressed value established by standard methods (constructing the MVR model based on all available samples) with the distribution of regressed values established from the held-out testing samples for each patient measurement ( $n = 174$ ) from each of the 4482 total measurements. The error associated with the standard use of MVR (all samples) as compared with our best bootstrap estimate (the mean of the regressed values across the bootstrapped set of MVR models) is assessed as

both the percentage error and with the  $z$  score (which assumes Gaussianity), which assesses the number of standard deviations the standard method is off from the bootstrap estimate. The average standard deviation of the regressed bootstrap distributions was also calculated.

In order to help illustrate an underlying reason why MVR models may be unstable, a sample size analysis was performed to assess the effect size associated with sex that the MVR models are trying to control for. The total dataset was randomly downsampled from 5% to 100% of the total dataset in steps of 1%, and at each downsampling step, 250 random selections of the current step's sampling percentage were included in a Cohen's  $d$  statistic calculation to assess how the sex effect changes with sample size.

### 3 | RESULTS

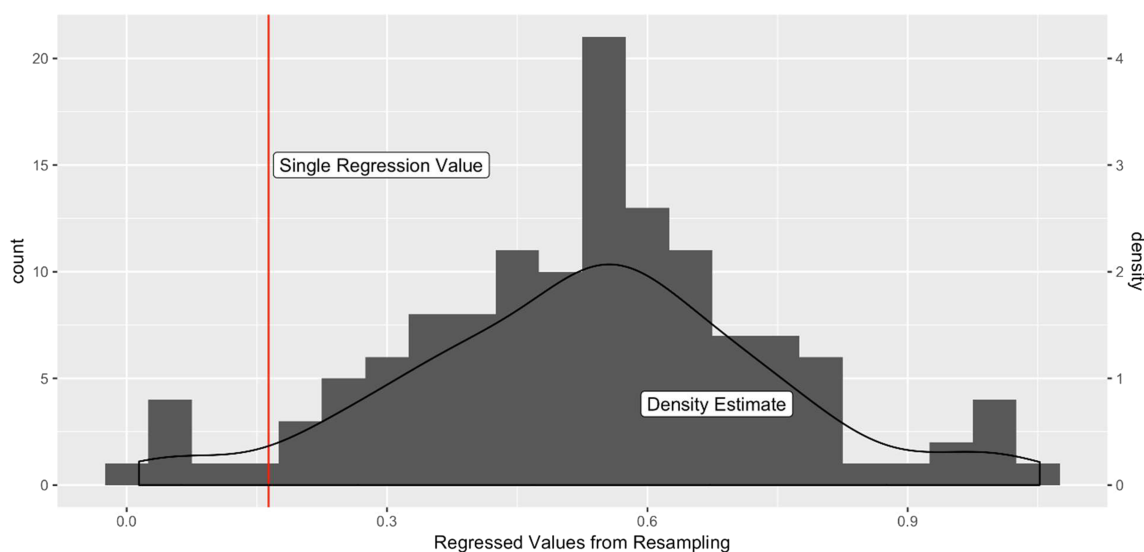
Results demonstrate that regressed measurements in this dataset deviate from the mean value of the sampling distribution created by the bootstrap (50/50) machine learning style validation, on average by 9.84%, which represents an average  $z$  score of 0.5 (which is also the number of standard deviations off of the bootstrap estimate on average). When switching to  $K$ -fold cross validation, the average error across all measurements was 9.7%. On average, the bootstrap distributions had a large standard deviation of 0.639. For example, total brain volume (from segmentation) extracted by FreeSurfer (Fischl, 2012) is a very common measurement considered in literature studies. Figure 1 provides the hold-out bootstrap distribution of regressed values for the first sample in our dataset on the total brain volume measurement for

illustrative purposes, to demonstrate what an example distribution might look like. Note the large width (high standard deviation) of the resultant sample distribution and the large error associated with the regressed value from the standard MVR model that included this sample in model construction (see vertical red line). Table 1 provides a listing of those measurements that exhibited the largest error on average across all samples by the standard method relative to the bootstrap machine learning validation method along with the  $z$  score and the standard deviation of the sampling distribution. Table 2 provides a listing of those measurements that exhibited the smallest error on average across all samples by the standard method relative to the bootstrap machine learning validation method along with the  $z$  score and the standard deviation of the sampling distribution.

Figure 2 is provided to demonstrate variability in the sex effect (as measured by Cohen's  $d$  statistic: the difference between the total brain volume means of the males and females divided by the standard deviation of the joint distribution) as a function of sample size. Note the large variability in effect size at low sample sizes and the lack of consistent and stabilized effect sizes at high sample sizes for this dataset. Figure 3 is provided to demonstrate the variability (as measured with the standard deviation) of sex effect sizes across sample sizes. Note the lack of stable low standard deviations at high sample sizes for this dataset.

### 4 | DISCUSSION

We have applied statistical machine learning validation techniques to the assessment of reliability and stability of



**FIGURE 1** An illustrative example of a hold-out validation bootstrap distribution of the regressed values associated with the first patient in the dataset's total brain volume measurement (Fischl, 2012)

**TABLE 1** Measurements exhibiting the largest discrepancies between the common use of multivariable regression and hold-out validation-based bootstrap estimate of multivariable regression

| MRI measurement  | Error % | Z score | Standard deviation |
|--|---------|---------|--------------------|
| Number of vertices on the left lateral occipital cortex    | 30.06   | 0.7952  | 0.9573             |
| Number of vertices on the left subcentral gyrus and sulcus | 28.34   | 0.7792  | 0.9538             |
| Surface area of the left rostral anterior cingulate        | 27.90   | 0.7589  | 0.9671             |
| Number of vertices on the left precentral cortex           | 27.78   | 0.7809  | 0.9632             |
| Surface area of the left lateral occipital cortex          | 27.29   | 0.7606  | 0.9441             |
| Surface area of the left planum temporale                  | 27.25   | 0.7670  | 0.9609             |
| Gaussian curvature of the precentral cortex                | 26.84   | 0.7800  | 0.9407             |
| Total gray matter volume                                   | 26.71   | 0.7981  | 0.9555             |

Abbreviation: MRI, magnetic resonance imaging.

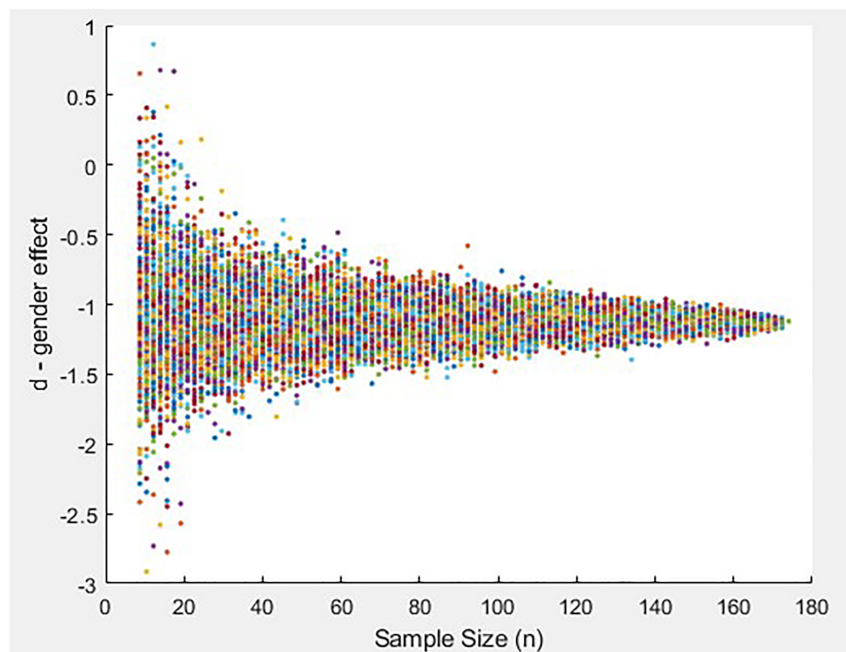
**TABLE 2** Measurements exhibiting the smallest discrepancies between the common use of multivariable regression and hold-out validation-based bootstrap estimate of multivariable regression

| MRI measurement   | Error % | Z score | Standard deviation |
|---|---------|---------|--------------------|
| Curvature index in the right anterior transverse collateral sulcus  | 5.26    | 0.3489  | 0.4520             |
| Standard deviation of the signal intensity of the left banks of the superior temporal sulcus white matter | 5.27    | 0.3204  | 0.4150             |
| Curvature index of the right inferior temporal cortex   | 5.28    | 0.3269  | 0.4280             |
| Left hemisphere positive surface integral   | 5.32    | 0.3378  | 0.4377             |
| Curvature index of the left inferior temporal sulcus  | 5.32    | 0.3440  | 0.4310             |
| Average thickness of the right parieto-occipital sulcus   | 5.33    | 0.3400  | 0.4439             |
| Number of vertices on the left medial occipito-temporal and lingual sulci                                 | 5.34    | 0.3458  | 0.4437             |
| Average curvature of the left Brodmann's Area 4a  | 5.34    | 0.3509  | 0.4595             |

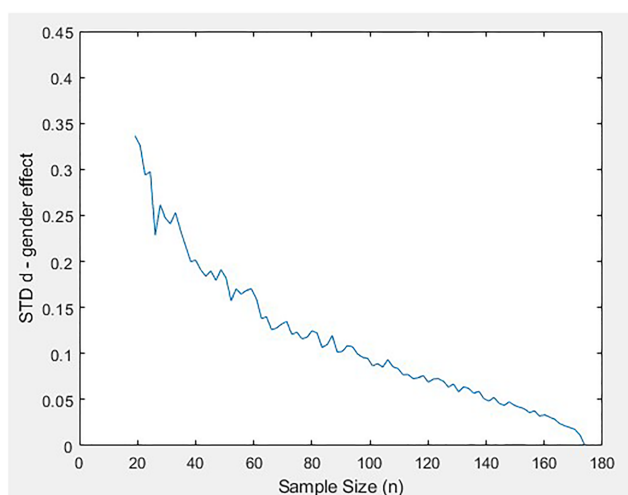
Abbreviation: MRI, magnetic resonance imaging.

MVR models. Results demonstrate that standard use of multivariable linear regression (construction of a single model without validation) can introduce large sources of error (up to 30.06%) in this publicly available dataset, although on average the error across all measurements was 9.84%. When switching to *K*-fold cross validation, the average error across all measurements was 9.7%, indicating that the introduction of error in MVR models occurs even when 90% of the samples in this dataset are devoted to model creation. Although this work demonstrates shortcomings of the use of MVR on a case study of brain MRI examinations from patients with schizophrenia, the analytic approach presented here can be applied to a

broad segment of neuroscientific studies that make use of MVR techniques to control for secondary effects in order to assess stability and reliability of the models produced. Without stable and reliable MVR models, attempts to control for (or adjust for) secondary effects may be associated with the introduction of new sources of error in an effort to avoid the error associated with secondary effects. Figure 1 demonstrates that a regressed measurement from a single patient can be very far off from the bootstrapped estimate produced with machine learning style hold-out validation. The wide standard deviation of the bootstrapped estimate distribution (see Figure 1) is another indication of MVR model instability. As such, we



**FIGURE 2** Cohen's  $d$  statistic assessed sex effect as a function of the fractional sample size of our dataset based on the total brain volume measurement



**FIGURE 3** The standard deviation of the distribution of Cohen's  $d$  statistic across sample sizes based on the total brain volume measurement

are not recommending that practitioners merely rely on the average of the bootstrapped estimate distribution (peak of distribution in Figure 1), nor the standard use of MVR (vertical line in Figure 1); instead, we are demonstrating that in applications where these two methods do not agree with each other, that MVR will be expected to be unreliable. In future studies, neuroscientists can reproduce the methods employed in this analysis to assess whether MVR is stable or unstable in any given application (with quantification of average error etc. that are associated with model stability) across all measurements under consideration. In applications where the standard

use of MVR does not agree with the bootstrapped distribution from hold-out validation, MVR is unstable to rely upon. The amount of error associated with the use of MVR that is tolerable in a given application is unknown a priori and the subject of future research. Ideally, an experiment would collect a sufficient number of samples for the hold-out bootstrap distribution of MVR regressed values (example Figure 1) to agree with the regressed values obtained from the standard use of MVR (example vertical line in Figure 1).

Figure 2 demonstrates high variability in the assessment of the effect size across varying sample sizes within this dataset. Note that the right side of the plot necessarily approaches the underlying effect size observable in the entire population included in this dataset. For the sample size to be sufficient for reliably establishing the sex effect, this plot would need to have a stable horizontal line of unchanging or nearly unchanging  $d$  values on the right side of the plot (i.e., at large percentages of the dataset included). Unfortunately, Figure 2 does not exhibit a stable horizontal line on the right side, which indicates that the sex effect has not stabilized at full sample size, implying that it is not possible to reliably establish the sex effect in this dataset, given the sample size constraints, and this feature is likely contributing to unstable MVR models. Figure 3 demonstrates the standard deviation of the assessed sex effect as a function of sample size. Similar to Figure 2, a stable sex effect would result in the right side of the plot exhibiting zero or near zero variability (i.e., the curve would fall to the bottom of the plotting area and stay there across high sample sizes). Because neither Figures 2 nor 3 stabilize at the highest sample sizes available in our

dataset, we expect that we have an insufficient sample size with which to properly model the sex effect in this study. Because many neuroscience studies are conducted using MRI at sample sizes similar or smaller to that used in this study, much existing research may have insufficient sample sizes in order to reliably perform MVR. This helps further motivate the existing trend of moving towards study designs with larger sample sizes to not only avoid reporting erroneous findings (through increased statistical power) but also to specifically aid in the creation of reliable MVR models. Fortunately, the validation techniques outlined in this manuscript can be replicated across neuroscience research studies in order to assess whether sufficient sample sizes are available to properly model the effects that they are trying to control for.

Tables 1 and 2 demonstrate that a wide variety of measurement types have both comparatively low and high levels of error associated with the standard MVR method relative to the machine learning style validation-based bootstrapped estimates. This implies that it would be difficult to predict beforehand which biomarker measurements will exhibit large amounts of error without completing the MVR bootstrapping technique employed in this analysis. It is expected that measurements with larger age and sex effects will be more negatively affected by the use of MVR at small sample sizes.

It is a given that a single multivariable analysis of a given sample will produce regression coefficients that will contain multiple sources of error, including randomly affected sampling error and bias. Bias can be problematic and potentially associated with measurement error or some other systematic error. When comparing the difference between the estimated regression coefficient from a sample to the “grand mean” of all sample coefficients, as we have done in this study, we are effectively relying upon the central limit theorem for assessing MVR stability. By comparing the error in this single estimate with the mean of multiple bootstraps, we are potentially able to unearth some systematic error where the single sample analysis will also contain the possibility of overfitting that is addressed in the repeated bootstrap sampling. Including larger sample sizes is widely expected to lead to a better estimate of the true regression coefficient mean. Future work will analyze a much larger dataset involving randomly varying the number of samples included in the analysis; then we will be able to quantify the variability of MVR introduced error across varying sample sizes for a given dataset.

## ACKNOWLEDGMENTS

This work was supported by the National Institutes of Health (grant numbers R01NS109475, R03NS091587, R21MH118739) to ET, Natural Sciences and Engineering

Research Council of Canada’s Canada Research Chairs grant (grant number 231266) to JL, a Canada Foundation for Innovation and Nova Scotia Research and Innovation Trust infrastructure grant (R0176004) to JL, a Natural Sciences and Engineering Research Council of Canada Discovery Grant to JL, a St. Francis Xavier University research startup grant to JL (grant number R0168020), and a St. Francis Xavier University UCR grant to JL.

## CONFLICT OF INTEREST

The authors have no conflict of interest to declare.

## ETHICS APPROVAL STATEMENT

This is an analysis of a publicly available dataset, for which ethics approval is not typically required. However, we also have approval for conducting retrospective analyses by Boston Children’s Hospital’s Institutional Review Board and ratified by St. Francis Xavier University.

## PATIENT CONSENT STATEMENT

Patient consent was provided as part of the collection of the publically accessed dataset.

## PERMISSION TO REPRODUCE MATERIALS FROM OTHER SOURCES

Not applicable.

## AUTHOR CONTRIBUTIONS

Conceived of analysis: JL. Supervision: JL, IG, ET. Manuscript authoring: JL, MJ, PK. Manuscript editing/approval: MN, JL, PT, ET. Software development and review: ER, MJ, JL, DB.

## DATA AVAILABILITY STATEMENT

The data that support the findings of this study are openly available: <https://coins.trendscenter.org/>.

## ORCID

Jacob Levman  <https://orcid.org/0000-0002-9604-3157>

Maxwell Jennings  <https://orcid.org/0000-0002-3285-5404>

Derek Berger  <https://orcid.org/0000-0003-4733-0624>

## REFERENCES

- Alexopolous, E. C. (2010). Introduction to multivariate regression analysis. *Hippokratia*, 14(Suppl 1), 23–28.
- Bzdok, D. (2017). Classical statistics and statistical learning in imaging neuroscience. *Frontiers in Neuroscience*, 11, 543. <https://doi.org/10.3389/fnins.2017.00543>
- Bzdok, D., & Ioannidis, J. P. A. (2019). Exploration, inference, and prediction in neuroscience and biomedicine. *Trends in Neurosciences*, 42(4), 251–262. <https://doi.org/10.1016/j.tins.2019.02.001>

- Bzdok, D., & Meyer-Lindenberg, A. (2018). Machine learning for precision psychiatry: Opportunities and challenges. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, 3(3), 223–230. <https://doi.org/10.1016/j.bpsc.2017.11.007>
- Chyzhyk, D., Graña, M., Öngür, D., & Shinn, A. K. (2015). Discrimination of schizophrenia auditory hallucinators by machine learning of resting-state functional MRI. *International Journal of Neural Systems*, 25(3), 1550007. <https://doi.org/10.1142/S0129065715500070>
- Davis-Stober, C. P., & Dana, J. (2014). Comparing the accuracy of experimental estimates to guessing: A new perspective on replication and the “crisis of confidence” in psychology. *Behavior Research Methods*, 46(1), 1–14. <https://doi.org/10.3758/s13428-013-0342-1>
- Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., Buckner, R. L., Dale, A. M., Maguire, R. P., Hyman, B. T., Albert, M. S., & Killiany, R. J. (2006). An automated labelling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, 31(3), 968–980. <https://doi.org/10.1016/j.neuroimage.2006.01.021>
- Fischl, B. (2012). FreeSurfer. *NeuroImage*, 62(2), 774–781. <https://doi.org/10.1016/j.neuroimage.2012.01.021>
- Fischl, B., & Dale, A. M. (2000). Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences of the United States of America*, 97(20), 11050–11055. <https://doi.org/10.1073/pnas.200033797>
- Fischl, B., van der Kouwe, A., Destrieux, C., Halgren, E., Ségonne, F., Salat, D. H., Busa, E., Seidman, L. J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., & Dale, A. M. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, 14(1), 11–22. <https://doi.org/10.1093/cercor/bhg087>
- Gollub, R., Shoemaker, J. M., King, M. D., White, T., Ehrlich, S., Sponheim, S. R., Clark, V. P., Turner, J. A., Mueller, B. A., Magnotta, V., O’Leary, D., Ho, B. C., Brauns, S., Manoach, D. S., Seidman, L., Bustillo, J. R., Lauriello, J., Bockholt, J., Lim, K. O., ... Andreasen, N. C. (2013). The MCIC collection: A shared repository of multi-modal, multi-site brain image data from a clinical investigation of schizophrenia. *Neuroinformatics*, 11(3), 367–388. <https://doi.org/10.1007/s12021-013-9184-3>
- Gould, I. C., Shepherd, A. M., Laurens, K. R., Cairns, M. J., Carr, V. J., & Green, M. J. (2014). Multivariate neuroanatomical classification of cognitive subtypes in schizophrenia: A support vector machine learning approach. *NeuroImage*, 6, 229–236. <https://doi.org/10.1016/j.nicl.2014.09.009>
- Greenstein, D., Malley, J. D., Weisinger, B., Clasen, L., & Gogtay, N. (2012). Using multivariate machine learning methods and structural MRI to classify childhood onset schizophrenia and healthy controls. *Frontiers in Psychiatry*, 3, 53. <https://doi.org/10.3389/fpsy.2012.00053>
- Iwabuchi, S. J., Liddle, P. F., & Palaniyappan, L. (2013). Clinical utility of machine-learning approaches in schizophrenia: Improving diagnostic confidence for translational neuroimaging. *Frontiers in Psychiatry*, 4, 95. <https://doi.org/10.3389/fpsy.2013.00095>
- Kambeitz, J., Kambeitz-Ilanovic, L., Leucht, S., Wood, S., Davatzikos, C., Malchow, B., Falkai, P., & Koutsouleris, N. (2015). Detecting neuroimaging biomarkers for schizophrenia: A meta-analysis of multivariate pattern recognition studies. *Neuropsychopharmacology*, 40, 1742–1751. <https://doi.org/10.1038/npp.2015.22>
- Nieuwenhuis, M., van Haren, N. E. M., Hulshoff Pol, H. E., Cahn, W., Kahn, R. S., & Schnack, H. G. (2012). Classification of schizophrenia patients and healthy controls from structural MRI scans in two large independent samples. *NeuroImage*, 61(3), 606–612. <https://doi.org/10.1016/j.neuroimage.2012.03.079>
- Schnack, H. G., Nieuwenhuis, M., van Haren, N. E. M., Abramovic, L., Scheewe, T. W., Brouwer, R. M., Hulshoff Pol, H. E., & Kahn, R. S. (2014). Can structural MRI aid in clinical classification? A machine learning study in two independent samples of patients with schizophrenia, bipolar disorder and healthy subjects. *NeuroImage*, 84, 299–306. <https://doi.org/10.1016/j.neuroimage.2013.08.053>
- Ségonne, F., Pacheco, J., & Fischl, B. (2007). Geometrically accurate topology-correction of cortical surfaces using nonseparating loops. *IEEE Transactions on Medical Imaging*, 26(4), 518–529. <https://doi.org/10.1109/TMI.2006.887364>
- Veronese, E., Castellani, U., Peruzzo, D., Bellani, M., & Brambilla, P. (2013). Machine learning approaches: From theory to application in schizophrenia. *Computational and Mathematical Methods in Medicine*, 2013, 867924. <https://doi.org/10.1155/2013/867924>
- Yang, H., Liu, J., Sui, J., Pearlson, G., & Calhoun, V. D. (2010). A hybrid machine learning method for fusing fMRI and genetic data: Combining both improves classification of schizophrenia. *Frontiers in Human Neuroscience*, 4, 192. <https://doi.org/10.3389/fnhum.2010.00192>
- Yarkoni, T., & Westfall, J. (2017). Choosing prediction over explanation in psychology: Lessons from machine learning. *Perspectives on Psychological Science*, 12(6), 1100–1122. <https://doi.org/10.1177/1745691617693393>

## SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

**How to cite this article:** Levman, J., Jennings, M., Kabaria, P., Rouse, E., Nangaku, M., Berger, D., Gondra, I., Takahashi, E., & Tyrrell, P. (2021). Hold-out validation for the assessment of stability and reliability of multivariable regression demonstrated with magnetic resonance imaging of patients with schizophrenia. *International Journal of Developmental Neuroscience*, 81(7), 655–662. <https://doi.org/10.1002/jdn.10144>