

Code error in Wang H, Avillach P. “Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning.”

Catriona Miller, Theo Portlock, Denis M Nyaga, Greg D Gamble, Justin M O'Sullivan

Submitted to: JMIR Medical Informatics
on: September 21, 2024

Disclaimer: © The authors. All rights reserved. This is a privileged document currently under peer-review/community review. Authors have provided JMIR Publications with an exclusive license to publish this preprint on its website for review purposes only. While the final peer-reviewed paper may be licensed under a CC BY license on publication, at this stage authors and publisher expressly prohibit redistribution of this draft paper other than for review purposes.

Table of Contents

Original Manuscript..... 4

Supplementary Files..... 7

 Multimedia Appendixes 8

 Multimedia Appendix 0..... 8

 Multimedia Appendix 0..... 8



Code error in Wang H, Avillach P. “Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning.”

Catriona Miller¹; Theo Portlock¹; Denis M Nyaga¹; Greg D Gamble¹; Justin M O'Sullivan¹

¹ Liggins Institute University of Auckland Auckland NZ

Corresponding Author:

Justin M O'Sullivan

Liggins Institute
University of Auckland
85 Park Road
Auckland
NZ

Abstract

n/a

(JMIR Preprints 21/09/2024:66556)

DOI: <https://doi.org/10.2196/preprints.66556>

Preprint Settings

1) Would you like to publish your submitted manuscript as preprint?

Please make my preprint PDF available to anyone at any time (recommended).

Please make my preprint PDF available only to logged-in users; I understand that my title and abstract will remain visible to all users.

Only make the preprint title and abstract visible.

✓ **No, I do not wish to publish my submitted manuscript as a preprint.**

2) If accepted for publication in a JMIR journal, would you like the PDF to be visible to the public?

✓ **Yes, please make my accepted manuscript PDF available to anyone at any time (Recommended).**

Yes, but please make my accepted manuscript PDF available only to logged-in users; I understand that the title and abstract will remain visible to all users.

Yes, but only make the title and abstract visible (see Important note, above). I understand that if I later pay to participate in a JMIR journal, I will be able to make my accepted manuscript PDF available to anyone.

No. Please do not make my accepted manuscript PDF available to anyone.

Original Manuscript

Code error in Wang H, Avillach P. “Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning.”

Catriona Miller¹, Theo Portlock¹, Denis M. Nyaga¹, Greg D. Gamble¹, Justin M. O’Sullivan¹

¹ Liggins Institute, University of Auckland, Auckland, New Zealand.

Wang & Avillach [1] developed a convolutional neural network (CNN)-based diagnostic classifier for ASD. After preprocessing the genomics data from the Simons Simplex Collection (SSC) [2], common variants that may be protective or pathogenic for autism were extracted based on a chi-square test. The authors then designed a CNN-based diagnostic classifier for ASD that had an accuracy and area under the receiver operating curve (AUROC) of 88% and 0.955 respectively. The predictor in Wang & Avillach [1] is currently considered the exemplar in the field, giving much more accurate predictions for autism than other studies [3].

However, upon inspecting the code and repeating the analyses, we contend that the method used is flawed and leads to an approximately 30% over-estimation of predictive ability.

Wang & Avillach [1] did not provide a GitHub link to the code that was used in their paper. However, code can be found in Dr Wang’s GitHub repository [4] that matches the results and figures in the manuscript.

The first error we identified in the code [4] is a disparity in the way variants were encoded, compared to the paper’s description [1]. The methods state ‘...0 denotes that both allele values are reference alleles, 1 represents one allele value is a reference allele and the other one is the alternate allele, and 2 represents both are alternate alleles.’ However, the code assigns missing alleles to 0, any reference alleles (i.e. 0/0, 0/1 or 1/0) to 1, and both alternate alleles to 2.

The second and more significant error, in terms of over-estimation, occurs in the data split for training and test sets. The methods state ‘...the SSC samples were partitioned into 2 sets based on random sampling of individuals into a training set (80%) and a hold-out test set (20%). There was no overlap of individuals across the 2 partitions’ [1]. However, the code uses different indexing methods for the test and training set. Because there appears to be no random seed in Wang & Avillach [1], we cannot reproduce the exact overlap as it occurred in the manuscript. However, simulations (n=100) using the code identified an average of 80% (sd = 1%) of the test dataset being represented in the training dataset.

We corrected error 1 and 2 in the code [5] and generated new models, using the 100 features that were identified in Wang & Avillach [1], and the genomics data from the SSC [2]. Simulations (n=100) with these models result in an AUROC of 0.61 (sd=0.02) and an accuracy of 60% (sd=0.02). This is 0.34 and 28% lower than the reported metrics, respectively, in Wang & Avillach [1].

The accuracy of the CNN-based diagnostic classifier for ASD presented in Wang & Avillach [1] is over-estimated by ~28%. We contend that Wang & Avillach [1] should be retracted according to the COPE guidelines. We also believe that an “expression of concern” should be published while retraction is being considered since the field is under active research and the citation rate is increasing.

Acknowledgements

Simons Simplex Collection project number 15286.1.1.

Data and code accessibility

Wang & Avillach 2001 code: <https://github.com/hms-dbmi/Haishuai>. Clone of the code and corrected code is in https://github.com/Catriona-Miller/SFARI_paper_clone

Simons Simplex Collection is accessible at <https://www.sfari.org/resource/simons-simplex-collection/>

References

- [1] H. Wang, P. Avillach, Diagnostic Classification and Prognostic Prediction Using Common Genetic Variants in Autism Spectrum Disorder: Genotype-Based Deep Learning, JMIR Med Inform 9 (2021). <https://doi.org/10.2196/24754>.

- [2] G.D. Fischbach, C. Lord, The Simons Simplex Collection: A Resource for Identification of Autism Genetic Risk Factors, *Neuron* 68 (2010) 192–195. <https://doi.org/10.1016/j.neuron.2010.10.006>.
- [3] S.A. Alowais, S.S. Alghamdi, N. Alsuhebany, T. Alqahtani, A.I. Alshaya, S.N. Almohareb, A. Aldairem, M. Alrashed, K. Bin Saleh, H.A. Badreldin, M.S. Al Yami, S. Al Harbi, A.M. Albekairy, Revolutionizing healthcare: the role of artificial intelligence in clinical practice, *BMC Med Educ* 23 (2023) 689. <https://doi.org/10.1186/s12909-023-04698-z>.
- [4] H. Wang, Haishuai - DeepAutism-ChiSquare, GitHub (2020). <https://github.com/hms-dbmi/Haishuai> (accessed September 12, 2024).
- [5] C. Miller, SFARI_paper_clone, GitHub (2024). https://github.com/Catriona-Miller/SFARI_paper_clone (accessed September 12, 2024).

Supplementary Files

Multimedia Appendixes

This slide deck presents a walk-through of the steps that were taken and the results of our attempt to reproduce the work of Wang & Avillach 2021.

URL: <http://asset.jmir.pub/assets/e17bb8f5c0fac490045d9923a459a49c.pptx>

Amended manuscript.

URL: <http://asset.jmir.pub/assets/416950f002ea7ab669caaf62da63fd26.docx>