NEWS

‘Science working as it should’: Autism blood signature study earns open post-publication review

BY **ISABEL RUEHL**

10 JANUARY 2023

Listen to this story:
https://www.spectrumnews.org/wp-content/uploads/2023/01/audio-d8aa1979-587f-4f9d-a098-82ab7432ae44-encodings.mp3

Like all scientists, Eric Courchesne is used to having his research scrutinized by peer reviewers prior to publication. But for his most recent **study**, which appeared in October in *Molecular Psychiatry*, peer review did not stop there. Shortly after publication, researchers took to online forums to critique the work — in real time, and in public.

“It was really valuable because it helps establish a communication that can quickly dispel misunderstandings by peers,” says **Courchesne**, professor of neurosciences at the University of California, San Diego.

In the study, Courchesne and his colleagues describe a machine-learning model that, based on gene expression levels in blood samples from 240 children aged 1 to 4 years old, could identify those who have an autism diagnosis with 80 to 86 percent accuracy. The team trained tens of thousands of different models on 73 percent of the samples, tested the best performers on the remaining 27 percent, and then used Bayesian averaging to collapse the top 742 models into a final “ensemble” version.

Days after the paper’s publication, an anonymous commenter questioned the study’s methodology in a post on **PubPeer**, an online platform for researchers to discuss journal publications. “The use of the same data for training and evaluation (testing) is not good practice and is almost guaranteed to result in overfitting and inflated estimates of performance,” the commenter asserted, under the alias *Cynanchum itremense.*

Spectrum reached out to independent experts for comment on the study and the anonymous post, after which one of those experts, **Dorothy Bishop**, emeritus professor of developmental neuropsychology at Oxford University in the United Kingdom, raised additional questions about the study on her own **personal webpage**, as well as on PubPeer. Her comments prompted replies from Courchesne and *Cynanchum itremense*.

“This is science working as it should,” Bishop told *Spectrum*, referring to open discourse in post-publication peer review. “People can raise questions about research, and the researchers engage with the comments, which is what happened here.”

“I didn’t feel that the answers addressed everything,” she added, “but I was glad that Dr. Courchesne replied, and his answers did clarify some points.”

“I enjoyed the post-publication peer-review comments and back-and-forth discussion,” Courchesne says. “I kind of wish that Dorothy had been one of the reviewers early on.”

Several independent experts told *Spectrum* that the methodological issue the anonymous critic raised may limit the generalizability of the study’s findings.

“They chose the models [for the ensemble] based on the test set performance, so that’s definitely a warning that there’s some kind of circular analysis,” says **Yanli Zhang-James**, associate professor of psychiatry and behavioral sciences at SUNY Upstate Medical University in Syracuse, New York.

Applying Bayesian averaging perpetuates the issue, in which the test dataset informs the models rather than being “held out” as a truly independent sample — a common problem in machine-learning analyses, Zhang-James says.

But Courchesne insists there was no circularity. “The original training was kept separate from the validation outside test set,” he says. “That enabled us to identify 742 models that were high performing and were demonstrated to be high performing in the test set.”

That part of the analysis “was very sound and comprehensive, and their results of those individual models were actually good already,” Zhang-James says.

Focusing on the performance of the final ensemble model misses the point of this study, Courchesne says. “The purpose of it, in my mind, was to test whether, using the same subjects, you get an improvement when you use the ensemble.”

And they did find an improvement: The ensemble model performed more accurately than most of the individual models. The team’s next step is to test whether the ensemble proves effective at identifying autistic children in an independent sample, Courchesne says.

Bishop's posts also launched a discussion on the reliability and reproducibility of the blood samples themselves.

"Gene expression levels could vary from occasion to occasion depending on time of day or what you'd eaten," she wrote on her blog. "I have no idea how important this might be, but it's not possible to evaluate in this paper, where measures come from a single blood sample."

Courchesne says his team did collect longitudinal samples for about 30 individual children from the training sample, separated by 9 to 24 months. About 91 percent of the 1,822 predictive models discovered via the full training set performed at the same high level in this longitudinal subset, Courchesne says.

Although the team wrote these results into an earlier unpublished version of the work, it was not included in the final publication.

"It wasn't a huge number," Courchesne says. "It's expensive, and hard for kids to come back." Still, the unpublished finding suggests the blood measures are reproducible, he says.

Plus, he adds, his team took other precautions: The same phlebotomist drew each child's blood and postponed the draw if the child had a fever, as immune activation could potentially alter gene expression levels, Courchesne says. The RNA was extracted from those samples exactly the same way by the same person, too.

"We were really uniform," he says.

Given those precautions, **Sek Won Kong**, associate professor of pediatrics at Harvard Medical School and a faculty member in the Computational Health Informatics Program at Boston Children's Hospital, says he is less concerned with issues of fluctuating gene expression levels due to time of day or diet — "although there must be a few genes that correlate with age, fluctuate over time and respond to environmental factors," he says.

Courchesne's team has already built up a dataset of more than 2,000 toddlers, and for their next study, they are testing the ensemble model's ability to discriminate between autism and language delay, he says.

Cite this article: https://doi.org/10.53053/PHME4244