Dear Editor

The authors thank the reviewers and the editors for their consideration, time, and thoughtful comments, which has undoubtedly improved this work. We summarize the changes, followed by detailed response to reviewer comments:

SUMMARY OF CHANGES:

1. **Typos corrected**
2. **Text added for clarifications in main and SI (in red)**
3. **Figure added to Supplementary text (SI-Fig 2)**

Reviewer: 1

This study describes a machine learning / inferred digital biomarkers approach for detecting or diagnosing the autism from individual diagnostic codes recorded during normal or frequent medical encounters / consultations. Based on probabilistic finite state automaton approach, the described risk estimator identifies children at risk with a corresponding area under the receiver operating characteristic curve (AUC) exceeding 80%from shortly after two years of age for either sex. The authors claim that the described framework and the autism co-morbid risk score has superior performance to the questionnaires-based screenings like the M-CHAT/F. My overall impression is that the study is well performed and the results are interesting. In what follows, I list a few issues
that require some further clarifications:

1) In the paper, the authors claim that "standard deep learning approaches do not yield sufficiently high predictive performance or statistical power...". Indeed, the results in SI Fig 3A do not seem promising, but the LSTM-M-truven is getting close to 80% similar to the SLD-M and SLD-F. One issue would to explain why the LSTM and other deep learning approaches fail and how do they compare in complexity to the proposed SLD approach.

RESPONSE:
There are two key and related issues in using deep learning approaches in this context. Both of these issues relate to the high sample complexity of standard neural network (NN) models (they tend to be data-hungry): 1) NN models generally have a relatively large number of parameters, and 2) these large number of trainable parameters require substantial amount of data to train. In our case, we indeed have a large number of control examples, but a relatively smaller number of positive examples (children who get diagnosed with autism eventually). The amount of training data, even with our large patient database, is not enough to train anything but the simplest of the NN architectures, and even then the number of parameters for such NN models is substantially larger compared to our architecture. This is shown in our example, where one of the LSTM models we investigate (the simplest one we tested) has 185,465 parameters, whereas our complete end to end pipeline has 7025+6719=13,744 trainable parameters (an order of magnitude less, where the first contribution is from inferred PFSA models, the second from the ensemble of decision trees in the final gradient boosting model).

Additionally, the number of parameters for deep learning models rapidly increase to millions or tens of millions as the models get more complicated. Also note that our framework is adaptive, and the network structures of the inferred PFSA models is inferred from data, and not fixed a priori; it produces more states where it needs, thus adapting the model resolution to more complex contexts, and obtaining significantly more compact yet effective representations. We must however add the caveat that with sufficient patience and experience it might be possible to devise a specific NN architecture with similar performance to our framework; however, in our approach such artful insight is unnecessary.

It also seems possible that NN models, which have no stochastic parameters or components, generalize poorly in this context, which is why the performance is relatively better in the Truven dataset, but worse in an independent database.

We have added this commentary in the Supplementary text, with a brief note added in the main text.

2) Can the authors comment on why the ASD related diseases (co-morbidities) are mutually independent? For example, if one has disease A, is he/she more likely to have disease B compared with the case he/she doesn't have disease A?

Diseases are not in general independent. Clearly complex dependencies exist between underlying pathobiological processes. These dependencies are indeed one of the reasons why our approach works. If all the diseases were independent, it would be impossible to estimate risk from co-morbidities. There is extensive literature of ASD comorbidities, which is easy to verify at the population level. This work shows that using our modeling we can leverage such patterns, even with ASD heterogeneity, to estimate individual risk.

3) If the co-morbidities are dependent, is it possible to eliminate dependence (data de-correlation)?

This is a nice idea. Although we are not sure how this relates to the problem at hand. Perhaps the reviewer is suggesting to decorrelate the signals and obtain predictive signatures which are independent of each other. While such an idea works quite elegantly in signal processing analyses in engineered systems, it is very difficult to make it work in the present context of sparse, noisy **categorical** diagnostic codes. We will explore this idea nonetheless in future work.

We have added this comment in the conclusion.

4**) In the RESULTS section, the authors state "Our approach produces a strictly superior PPV (exceeding M-CHAT/F PPV by at least 14%... around age of 26 months". My understanding is that this happens around 104 weeks but the paper indicates 112 weeks. While this requires a minor clarification, it would be useful to see a concrete comparison between the proposed and other methods at specific ages. Is the proposed method better for early on prediction or gets more confident later?**

26 months is 112-113 weeks (not all months have exactly 4 weeks, and a simple calendar lookup will confirm this, e.g. here https://www.datecalculator.org/months-to-weeks). Noted in main text.

To the other question, we don't have detailed baselines at all ages to compare to. But clearly our approach becomes much better and more confident with time, reaching nearly 90% AUC approaching 4 years, as illustrated in Fig. 2a.

5) The proposed prediction method is mainly based on mathematical modelling of the co-morbidities and machine learning classification techniques. Is it possible to combine some domain knowledge (e.g., at what age ASD is the most common; influence of inheritance disorders etc.) to improve the prediction performance?

Quite certainly that should be the case. However, in this paper we intended to focus on what is possible with data that will be readily available at the point of care. More detailed information such as genetic markers, or familial data might make the predictions better, but takes us away from our intended aim of developing a universal screening tool. In future we will explore these modifications in more details, particularly the predictive boost by adding maternal codes, and also genetic information on the children to begin with. We want to add here that our analyses DOES use at least some domain knowledge: 1) we focus on risk estimation at around 2-3 years, since it is at this age diagnosis is the most effective, 2) we demonstrate our performance conditioned on M-CHAT/F scores, which is the current standard tool for ASD screening. Have commented in the Conclusion.

6) Is the prediction dynamic and time-sensitive? For example, if we are recording a child's ASD co-morbidities record and the child doesn't show a sign of ASD at some age, can we determine he/she is healthy and terminate the observation and prediction? In a different case, if he/she shows some sign of ASD, how long should we proceed monitoring his/her behavior and healthy condition to make the prediction?

The ACoR is demonstrably time-aware. This is shown in multiple panels in Fig 2B and 2E. We also see that a strongly negative result at around 2 years will not abruptly switch and become positive later if sufficient codes have been observed before. Every prediction we make has an uncertainty bound on it, allowing us to make probabilistic statements with quantified uncertainty on the eventual status.

We have clarified this in the caption of Fig. 2 and in the main text.

7) There are also a few minor issues like grammar misspellings or typos:
- “comprising de-identified” probably should read “comprising of de-identified” on page 3
- “Each of these inferred models in a Probabilistic Finite State Automaton” missing verb or some other information is missing

- “correspond to the control” in caption of Fig 1
- “set feasible set” on page 9
- “Th ACoR is free”

Fixed in manuscript (except the first one, which is actually grammatically correct: “comprising de-identified" is the correct phrase.

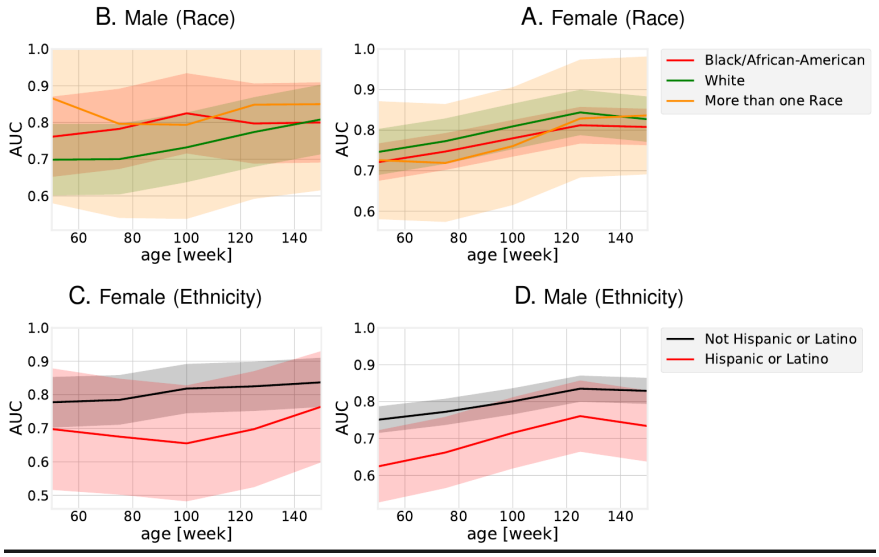
Reviewer: 2

This manuscript reports the estimation of a complex and novel machine learning architecture to a large and diverse sample of electronic health record data, improving on the positive predictive value of state-of-the-art screening markers for Autism Spectrum Disorder (ASD) of equal scalability. The

results yield novel insights that may be reasonably be expected to trigger significant breakthroughs in treatment of this condition, where core symptoms are unaddressed by any existing pharmacotherapy. Furthermore, there is limited evidence of efficacy of any modality, behavioral or pharmaceutical or otherwise, from adequately-controlled randomized controlled trials for these same symptoms. As a new reviewer of this manuscript, my sole critique of the paper relates to the issue addressed chiefly in Si Fig 1, which is of significant public interest that it be more adequately addressed in the main text. In particular, is the ACOR really free from bias due to systematic under-diagnosis in diverse populations?

The authors make this important assertion, but it is not clear whether or to what extent it holds. The alternative hypothesis - that ACOR inherits the systemic biases that already exist in diagnostic procedures, and may even compound them, even if it improves diagnostic accuracy overall (relative to existing procedures) - needs to be articulated in the main text. This issue should not prevent the important results of this work to be communicated to the global audience of Science Advances, but simply needs to be more focally acknowledged in the main text so that the results are not used in ways that increase existing disparities in diagnosis and access to care.

This is an extremely important point. We thank the reviewer for pointing this out. We have explicitly compared this now with data from the UCM database which has demographic information (the Truven database does not), and results are added in the SI text, with note added to the main text. We find that while children of Hispanic/Latino ethnic background have a lower AUC, the differences are not significant. Similarly, comparing between white, African-American and multi-racial racial backgrounds, we found no significant differences in the performance of our algorithms. Other races or ethnicities could not be compared due to lack of sufficient data, and will be investigated in future work.



The authors should be commended for making their model available and reusable under non-restrictive terms, to maximize the surely global impact of this work.

Minor Typos in main text:
"Each of these inferred models in a Probabilistic Finite State Automaton (PFSA)" should probably read "Each of these inferred models is a Probabilistic Finite State Automaton (PFSA)".

"with only state of the art machine learning the predictive performance is significantly worse" should probably read "with only prior state-of-the-art machine learning the predictive performance is significantly worse"

"we identify the set feasible set of conditional choices in a four-dimensional decision space" should probbaly read "we identify the feasible set of conditional choices in a four-dimensional decision space"

"Th ACoR is free from aforementioned biases, and yet significantly outperforms the tools in current practice." should read "The ACoR is free from aforementioned biases, and yet significantly outperforms the tools in current practice." (but see my note above)

Fixed in text