Revision Comments from the June 8, 2020 submission of “Modeling *Plasmodium falciparum* Diagnostic Test Sensitivity using Machine Learning with Histidine-Rich Protein 2 Variants” by Ford et al.

(PONE-D-20-17545)

|  |  |
| --- | --- |
| **Reviewer Comment** | **Action Taken** |
| The authors make very broad claims regarding their approach toppling the past 2x7 model by only relying on Pf parasites from a single site in far western Ethiopia. While their framework for modeling and identifying important motifs of PfHRP2 is valid, they should be careful to understand they are only describing the Pf parasites in their study, and previous work by groups in their References (and others) have made it a point to investigate Pf genetics from isolates around the globe. Specifically, the original Baker report and follow-up investigations emphasized the type 2 and 7 repeats as they were so reliably found among nearly all global isolates. The Discussion in the current draft is very lacking and contains only a single reference. The authors need to emphasize what their approach identifies about this particular Pf population at this site in Ethiopia and how these ML approaches would be useful for meta-analyses of global Pf sequences. | Our original submission was quite clear that these results are for our samples collected in a specific area of Ethiopia and that more work needs to be done with data from other regions. However, we have added additional verbiage on this, focusing on global applicability of the methodology rather than the global applicability of these Ethiopian results. |
| A mention of monoclonal antibody binding to PfHRP2 epitopes is completely missing from this report. The Discussion would be a good place to expand on what other groups have learned about the major and minor epitopes for mAb binding to PfHRP2 and implications for their ML findings for better identifying RDT targets. | Added to discussion section of the manuscript. |
| As currently written, the report has an imposing 14 elements (4 Fig and 10 Tables). Many of these elements, like Table 1-5, would be better suited for Supporting Information for a clearer report. | No action taken. Will be fixed in the copyediting process. |
| Abstract: ‘to characterize malaria infections’ is very vague in this context. Provide an explanation of what type of characterization. | Addressed in manuscript. |
| Abstract: ‘to assert P. falciparum infection’, again a vague context here. Perhaps replace with ‘to describe this entire protein’ or similar. | Addressed in manuscript. |
| Intro: ‘infections are can treated using’, remove ‘can’? | Fixed in manuscript. |
| Intro: artemisinins and ACTs are just one form of anti-malarial, and not used in every country. Please rephrase to be inclusive of all antimalarials here. | Fixed in manuscript. |
| Intro: ‘Unfortunately, the efficacy of RDTs and artemisinin treatment are both waning’. In some very specific settings, this has been seen, but is not generalizable to the whole malaria-endemic world. Please rephrase to emphasize the very specific locations where deletions or K13 mutations observed. | Fixed in manuscript. Added verbiage on K13 protein mutations. |
| Intro: ‘Our purpose is to use large datasets and machine learning methods to address the shortcomings in malaria diagnosis.’ Very open-ended here. Current RDTs do a very good job of diagnosing malaria almost everywhere. Rephrase this sentence to give the specific objective of your study in context of characterizing HRP2. | Fixed in manuscript. |
| Intro: ‘We apply this model interpretability here to identify which types of histidine-rich repeats in PfHRP2 are most indicative of malaria infection.’ I think the concept here is understanding which PfHRP2 repeats are most seen during active Pf infections to better design RDTs(?) If any repeats are present, this is 100% indicative of Pf infection. | Fixed in manuscript. This study looks at the discrepancy between existing rapid diagnostic tests versus PCR-based lab tests. |
| Although the authors claim that they used a dataset of 406 PfHRP2 sequences, only 102 sequences were successfully generated. | Fixed/clarified in manuscript. We collected 406 samples, of which 102 successfully sequenced. |
| A key methodological issue concerns on how the quality control of sequencing data was performed. How many times each sample has been sequenced? What algorithm was used to check the quality of sequences? Did the authors combine the forward and reverse DNA strands to obtain one consensus sequence for each sample? Because many sequences presented stop codons, it seems that there were some of them with sequencing errors. | Added information in manuscript. |
| In sequence translation step, why did the authors not generate one consensus sequence and analyzed this unique sequence per sample? Combining the forward and reverse strands from sequenced PCR products (and to repeat the sequencing from a PCR product independently) is important to eliminate sequencing errors mainly when a proofreading enzyme is not used. If there was any concern about multiple-clone infections, other approaches should be applied, e.g. cloning of PCR products and performing MSP-1/-2 or microsatellite typing. | Added information in manuscript. |
| What are the main characteristics of the study patients? Were all of them symptomatic? I presume that all cases were symptomatic since they were attended in healthy clinics. What about the study area? When were samples collected? | Added information in manuscript. |
| What does it mean "1" represents a positive case of malaria and "2" represents a negative case of malaria for the dependent variable? Presuming that all cases were symptomatic and carried the parasite (because only who carries the parasite could be sequenced for HRP2), how the authors defined a negative case of malaria. Maybe it refers to the RDT result, but it is completely unclear. This raises another question how the sensitivity of RDTs was assessed (no bands, weak bands?). Since it is a fundamental issue, I can conclude that the results do not support the conclusions of the study. | Added information to clarify in the manuscript. This was originally shown in Figure 1. |
| What are the main characteristics of the study patients? Were all of them symptomatic? I presume that all cases were symptomatic since they were attended in healthy clinics. What about the study area? When were samples collected? | Added information in manuscript. |
| Each Table and Figure were presented out of order. | No action taken. Will be fixed in the copyediting process. |
| In Introduction, Malaria does not infect. Plasmodium infects and the disease cause or affect. | Fixed phrasing in manuscript. |
| In Repeat Type Prevalence (3.1), the results are mixed with discussion. | No action taken. Unclear what is meant by the reviewer. |
| In Fig. 4, describe the localities of the study in the Materials and Methods section. | No action taken. Unclear what is meant by the reviewer. |