nature portfolio

Peer Review File

Early assessment of antibodies decline in Chagas patients following treatment using a serological multiplex immunoassay

Corresponding Author: Dr MAAN ZREIN

This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
In the manuscript by Ursula Saade et al., entitled “Early prediction of parasitological cure in Chagas patients: a new serological multiplex immunoassay,” the authors evaluate anti-T. cruzi antibody dynamics to a set of antigens on the commercial MultiCruzi assay (InfYnity Biomarkers) in clinical trial specimens from the BENDITA trial to evaluate prediction of parasitological cure. There is a significant need for biomarkers that are more responsive to efficacious treatment considering the current outcomes used in drug trials either use PCR in peripheral blood, which is only positive in a fraction of chronic Chagas disease cases, or more recently a 20% decrease in signal from commercial ELISAs, which are not meant to be quantitative.

The MultiCruzi assay is unique in that multiple antigens printed on discrete dots within a microplate well capture antibodies from serum/plasma and generates a colorimetric reading of the individuals dots by image analysis. These colorimetric readings are normally applied to a cutoff, however, this paper employs an innovative method of serial dilutions to generate a sigmoidal curve (signal vs. titer), then generates a statistic of the halfway point along the linear portion of the dilutional curve (DF50) that can be compared between treatment timepoints from the trial samples (prior to treatment, 6 and 12 months). This is more efficient than end-point titers for each of the 15 antigens.

The findings of the study are positive, demonstrating significantly negative slopes of DF50 for certain antigens when plotted against time at 6 and 12 months in the treatment arm compared to the placebo arm. Furthermore, when compared to the previous method of 20% decrease in commercial ELISA signal, the signal change from MultiCruzi that the authors associate with efficacy of treatment was seen in a larger proportion of individuals in the treatment arms. The authors therefore conclude that the MultiCruzi assay is a better predictor of parasitological cure. These findings are consistent with previous publications of MultiCruzi in pediatric patients of the BENDITA trial, which are able to achieve seroreversion (to negative) (LJ Medina et al., 2021).

These results are promising and analytically demonstrate a robust method for evaluating dynamic antibody changes during treatment. The difficultly with this paper and discussion is that the authors seem to “jump the gate” when claiming this assay is a better predictor of parasitological cure. This is concerning when viewed through the lens that the first and corresponding authors are from the company that manufacturers this assay. As the introduction states, there are no tests for parasitological cure as PCR in peripheral blood only measures parasitemia (not potential intracellular reservoirs) and seroreversion (to negative) in adults only happens over years to decades, which is not feasibly demonstrated in the duration of this clinical trial. The outcome of parasitological cure in the adult specimen set referenced (13) are determined by PCR over 12 months. By this logic, this assay cannot infer predicted parasitological cure as this manuscript claims, only that this analytical method of interpreting antibody decreases to specific antigens in the MultiCruzi assay is correlated with treatment and more responsive than conventional serology. This study is also lacking some key validation experiments, including defining precision of the DF50 in technical replicates high, medium, and low signal specimens to understand the variance that may occur within and between runs.

Lastly, this manuscript only documents the specific antigens used in the supplementary materials and does not discuss the findings in terms of host-pathogen interactions. This would be a fascinating biological study that deserves some mention,

although I admit may be too broad for the focus of this paper.

Overall, this study is innovative with exciting findings that are likely applicable to further clinical trials, but more in terms of analytical laboratory methods for assessing T. cruzi antibody dynamics versus actually demonstrating a method with adequate validation for assuring parasitological cure.

Minor Comments:
Line 52: This line makes it seem as if sudden death is the most likely outcome. For clarification, I would either remove this or add a few more words about the spectrum of diseases caused by T. cruzi infection.

Line 70: The use of "erratic" implies that the parasitemia has major fluctuations over time. It would be more appropriate to say "unreliable as a surrogate for the overall parasite load in other compartments and tissues."

Line 72: Saying PCR is inappropriate for assessing drug efficacy is a bold claim when it was used as an outcome measure in the BENDITA trial for which these specimens were generated. This is additionally concerning given the commercial interests of the authors. PCR is an analytically sensitive and specific method of direct parasite detection, the flaw is in the low parasitemia due to the biology of parasitic infection in humans. The authors note that the flaw in study design is that PCR detection in blood is not present in all people, but this does not make it inappropriate as for assessing drug efficacy for which PCR positivity is an inclusion criteria.

Line 92: Should have a reference if including a statistic.

Line 108. These references are formatted differently for some reason.

Line 177: Figure 1a shows images of the antigens dotted on the well plate. In some high-intensity dots there is a trail or bleeding of color generated that drifts away from the dot and, in some instances, overlaps with other antigens or adds a tint to the whole well. Is this reviewed, corrected for or controlled in any way?

Lines 179, 181: DF50 is formatted differently than previously in the paper.

Reviewer #2

(Remarks to the Author)

The article represents a new confirmation of the usefulness of a new serological kit for the follow-up of patients with Chagas disease who have received etiological treatment.
The data obtained seem very interesting because, unlike current commercial kits, a more rapid drop in antibody levels is observed. It is well written, and the graphs show the results obtained in a very visual way. In any case, I think that before being so emphatic with the results, some extra analysis would be needed.
To evaluate such a test, the authors have used samples from patients included in the Bendita clinical trial, which evaluated different therapeutic schemes. Interestingly, all patients who have been exposed to the drug have had a response measured by the level of antibodies (more or less early depending on the treatment arm). Moreover, without a doubt this may represent a therapeutic effect of the medication.
In a second step, the authors reinterpret the data, and depending on the proportion of antigens that present a greater drop in this response, they venture to create a “cure” score.
What underlies healing is the presence or not of a viable parasite in the tissue, which on the one hand activates an immune response (from which healing can be inferred indirectly from its quantification) as well as the inflammatory trigger that triggers tissue damage. and future morbidity and mortality. The methods used to date to measure therapeutic efficacy are the detection of DNA in peripheral blood using molecular biology (PCR). As the authors comment, this is done with the intention that if the treated patient has a positive post-treatment PCR, the patient is considered to have failed, since it is accepted that this DNA comes from an active infection.
Therefore, this method can provide an idea of what the kinetics of antibodies after treatment represent, but to attribute to them a “usefulness” as biomarkers of cure, it is necessary to associate these antibody levels with clinical/prognostic data (not feasible within of the time frame of any study) or the presence/absence of the parasite. The authors have an excellent opportunity that I believe they should not miss, which would be to analyze what the authors define as the risk of cure with the PCR results. This is essential since we can have the paradox of having patients with decreasing antibodies with sustained detection of parasite in the blood.
In short, in order to understand this technique as a biomarker of cure, an analysis is needed that more or less indirectly confirms this outcome, and in this case, it would be the PCR.

Reviewer #3

(Remarks to the Author)
Phase 2 trials of anti-parasitic interventions for chronic Chagas disease in adults are a major hurdle. Chagas research is hampered by the lack of understanding of the disease pathogenesis and the factors associated with disease progression. A specific problem for phase 2 trials is quantifying parasitic cure. The parasite densities in blood in chronically infected adults are very low (usually less than 1 parasite per ml). These densities are only just detectable by qPCR targeting the minicircle

or satellite DNA. Drugs for Neglected Diseases initiative (DNDi) has done a series of phase 2 studies which have used "sustained parasite negativity" as the primary endpoint. The use of this endpoint is justified by the ability to detect very low densities using PCR (PCR can in theory detect 1 parasite in 5 to 10ml of blood). Sustained parasite negativity is defined as undetectable parasites in blood from serial PCRs done over 1 year of follow-up. The problem with this endpoint is that not detecting parasites (a series of negative PCRs) does not mean absence of parasitaemia due to the very low densities. It is not possible to completely rule out the possibility of continued infection. There is a need for better endpoints to determine full cure, compare drugs, and optimise doses. In this paper, the authors explore the use of serological endpoints to determine cure.

MultiCruzi is a multiplexed antibody assay for T. cruzi containing 15 antigens printed onto 96-well plates. From serial dilutions, it is possible to derive a quantitative output for each of the 15 antigens. The authors have used samples from a phase 2 placebo-controlled randomised trial of different benznidazole regimens, some in combination with fosravuconazole (the BENDITA trial) to assess the predictive value of the MultiCruzi assay for "serological cure". The authors derive a 50% dilution factor (DF50: the dilution estimated to give 50% of the maximal signal intensity) for each antigen. This is estimated from 3 serial dilutions (1/50, 1/400, 1/3200). Changes in DF50 over time are compared between the randomised arms under a linear model. They conclude that samples taken 6 months after treatment initiation can predict "treatment efficacy", and this performance is much better than standard ELISA tests. They propose using the MultiCruzi output to define serological primary endpoints for "proof-of-concept" (phase 2) trials in chronic Chagas disease.

The data contained in this paper are very interesting and novel. Finding serological markers of cure that could be measured shortly after end of treatment would be a major advance for Chagas disease drug development. However, the statistical analyses presented are not appropriate and the conclusions are not warranted.

Major comments

1. In the abstract, the authors suggest including MultiCruzi in the primary endpoint for proof-of-concept trials (currently this is PCR). The authors do not carry out a formal comparison between Multicruzi and qPCR. This suggestion is not justified based on the presented analyses.
2. The primary endpoint in the original BENDITA trial was sustained PCR negativity. Although PCR is an imperfect endpoint, it has high specificity. Therefore, a positive PCR result during follow-up strongly indicates that the individual still has replicating parasites in their body. The analysis presented here does not incorporate any PCR outcome data. No reason is given for this omission. The serial PCR data were not used to calibrate the observed changes in antigen reactivity. By considering the PCR outcomes in the trial, it is clear that the derived thresholds for the antigen reactivities are mis-calibrated. The proportion of patients predicted to be cured is too high. Table 1 shows the proportion of patients with "predicted cure" using the derived serological threshold from the analysis. In this table, it is estimated that 10 out of 30 patients given placebos were "cured" over one year follow-up. But this contradicts the PCR data in the original trial publication in Lancet Infectious Diseases: only 1 individual out of 30 had sustained PCR negativity over 12 months. This implies that 9/10 of the patients with predicted serological cure had in fact positive PCR results during follow-up: i.e. they were not cured.
3. There is value in comparing the longitudinal antibody patterns between the placebo and treatment groups. This is useful as a descriptive analysis. It is encouraging that there is more discriminative signal for the MultiCruzi assay compared to conventional serology. But (as mentioned in point 1) to show added value relative to PCR, and to argue that MultiCruzi should be used as a primary trial endpoint (instead of PCR), a lot more work needs to be done. There is already a huge and easily detectable difference in terms of PCR read-out between no treatment and relatively short treatment courses of benznidazole. The BENDITA study enrolled patients on the basis of a positive PCR at screening. 29 out of 30 patients randomised to placebo were consistently PCR positive throughout follow-up, whereas few of the benznidazole treated patients had positive PCRs during follow-up and those who did were generally positive at only a single timepoint (possibly indicating lower parasite densities much less that the lower limit of detection). Thus, PCR is very good at discriminating between treatment versus no treatment. The experience with fosravuconazole also indicates that PCR is good at determining whether a treatment does not work. What is less clear, however, is whether PCR can differentiate between slightly different treatments. One main output of the BENDITA trial was that 2 weeks of benznidazole appears to have similar efficacy as 8 weeks. If 2 weeks of treatment were sufficient, this would be a huge gain, as most side effects to benznidazole start in week 3, thus substantially improving adherence. I do not see how the analysis of the MultiCruzi data helps us to understand what the differences between these two treatment options are (if any).
4. The presentation of the statistical analyses is inadequate. We thank the authors for sending their code and data. However, the code was written in SAS and was contained in a Word document. We note that SAS is proprietary software, and so we could not re-run the code. However, we did some exploratory analysis of the data in R. After looking at the data, we question some of the authors conclusions. Many of the DF50 estimates appear truncated, with values of either 0.1 (lowest value) or 6400 (highest value). We do not have the raw reactivity data but can guess that in these instances, the sigmoid fit "failed" (all observed reactivities were high, or all were low). For antigens 110 and 134, >90% of the values are truncated! We take this to mean that the dilutions used were not appropriate for these antigens. Antigen 99 has 60% of values truncated and antigens 36 and 101 have ~33% truncated. So out of 15 antigens, 5 provide little to no information (you cannot estimate change over time using a truncated datapoint). There are only 6/15 antigens which have less than 10% truncation. These issues are not reported in the paper. Nor does the statistical model take truncation into account. In summary the statistical analysis and presentation of the data is inappropriate.

Minor comments

1. Figure 1 shows well images and normalised signal intensities for two patients in the BENDITA trial at baseline, 6 months, and 12 months after treatment initiation. Are the actual data points being shown in panels b & c? It looks like they fit the

sigmoid line perfectly, which makes me think they are not the actual data but predictions under the sigmoid model. The DF50 line should be vertical instead of horizontal (the units of DF50 are dilutions, not the signal intensity). The image in Figure 1 is useful and seems to suggest that the intensity measurement might be affected by “smearing” (not sure what the technical word for this is!). How is this captured in the digital data? More generally, how much noise is there (there are duplicate spots for each antibody which could help answer this question).
2. As a general point, none of the Figures appear to show any raw data. Because no raw data are presented, it is very difficult to get a sense of the variability of the measurement, the goodness of fit and the true differences between the intervention arms. I think code and de-identified data should be made openly available for reproducibility. Code should be written in open source software (eg R or Python).
3. On lines 68-72 the authors state: “However, T. cruzi is an intracellular parasite with cyclic and low parasitemia while it resides in the organs and during the indeterminate form of chronic Chagas disease. Its presence in the blood stream is thus erratic and does not represent the overall parasite load in other compartments and tissues. Therefore, the sensitivity and specificity of PCR methods are highly variable, making PCR inappropriate for assessing drug efficacy”.
qPCR is a well-established marker of drug failure. This is because PCR can detect very low numbers of circulating parasites in blood. It is not the specificity of PCR which is variable, it is the sensitivity. This sentence should be clarified and is not justified as PCR has not been compared with MultiCruzi.
4. Most of the results are given in terms of significance testing. The emphasis on p-values is inappropriate. It is not very useful to know that the difference in reactivity changes over time between two groups is significant or not. This is primarily driven by the sample size. What we would like to know are the quantitative differences. For example, the statement “Slopes of all the treatment groups were significantly more negative than the slope of the placebo group ($p < 0.0001$), indicating that the antibodies decline more rapidly in treated patients than in patients administered the placebo.” The p-value is not useful here. No magnitude of effect is given. In addition most of the results are presented without confidence intervals and with superfluous precision in the estimates (for example a slope coefficient of -0.02766: do we need 5 decimals points?). I would suggest standardising the reactivity measurements so that the mean value at baseline for each antigen is 100. Then all results could be presented as a % change over time.
5. There seems to be arbitrary post hoc reporting for particular antigens. Why is antigen #11 special (main text) or antigen #3 (Figure 1)? There does not seem to be any formal comparison of the 15 antigens, are there any which look better than others?
6. Use of the Youden index implies that the same weight is being given to the sensitivity of a cut-off value as to the specificity of the cut-off. But should that be the case here? Is it not more important to identify treatment failures?
7. Why not use a probabilistic model of cure? The set of rules proposed appear to be completely arbitrary and it is very unclear what the authors are trying to optimise for. As noted above, the predictions are known to be wrong (this predicts that 1/3 placebo patients are cured which contradicts known PCR data).
8. The Introduction is very long, it could be shortened considerably.
9. Please clarify the objective of the study. The introduction section (Line 123) and the discussion section state different objectives.
10. Line 138 in discussion mentions a variability in the ELISA tests (used in the BENDITA trial) that was not mentioned in results or analysis. This should be clarified.
11. Line 111-112: the authors should state that there is actually a need for a reliable biomarker of early treatment response of any type. It does not necessarily have to be a serologic biomarker.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Version 1:

Reviewer comments:

Reviewer #1

(Remarks to the Author)
The comments are sufficiently addressed.

(Remarks on code availability)

Reviewer #2

(Remarks to the Author)
My comments have been well argued and responded. Their publication may be very interesting for the scientific community.

(Remarks on code availability)

Reviewer #3

(Remarks to the Author)
The authors have made some changes to the paper and have added an analysis using the PCR data (primary endpoint in the BENDITA trial).
I find some of the additional Figures very interesting and useful, especially Fig 5 which shows the effect of the threshold on the estimated “response to treatment”.
However, I remain of the opinion that this work is of insufficient quality for publication as currently presented. My main issues are:
• Data are not provided in an appropriate format resulting in work that is not reproducible. This is especially problematic because the main authors are from the company that make the assay (and therefore have a clear conflict of interest). Transparency in data is paramount.
• The comparison between PCR and serology could be improved substantially. This is a classic “latent variable” problem. We can posit a true (unknown) “cured” versus “not cured” latent outcome for each individual in the trial. Both serology and PCR have different operating characteristics, notably +PCR has very high specificity for determining “not cured” and seroreversion has very high specificity for determining “cured”. A latent variable model would allow for an objective estimation of the operating characteristics of both these tests conditional on the model assumptions. This would be a very interesting and highly publication worthy exercise. As currently written it seems like the authors are presenting highly ad hoc rules for determining “response to treatment” using DF50 units which are difficult to interpret. The section on PCR is really difficult to read.
• The authors completely ignore the key question around how many of the 15 antigens on the MultiCruzi are actually needed. As noted before, 5/15 antigens provide very little information (mostly truncated values). But how many of the remaining 10 are useful? Could all this just be done with 1 or 2 antigens?
• I question the statement “dilutions of 1/50, 1/400, and 1/3200 were found to be optimal for all fifteen antibodies assessed with the MultiCruzi assay”. 5 of the 15 provide very little information: ie dilutions are not optimal?
• Most of the paper is still framed in terms of significance testing which is not very useful or informative. The authors’ response to my previous comment on this seems to miss the point.

(Remarks on code availability)
Not appropriate. They are using proprietary software and raw data are not made available. The code is given in the Word document. This is not of sufficiently high standard for reproducibility.

Reviewer #4

(Remarks to the Author)
I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

(Remarks on code availability)

Version 2:

Reviewer comments:

Reviewer #1

(Remarks to the Author)

(Remarks on code availability)

Nature Communications

Early prediction of parasitological cure in Chagas patients: a new serological multiplex immunoassay

Point-by-point rebuttal

Reviewer #1 (Remarks to the Author):

In the manuscript by Ursula Saade et al., entitled “Early prediction of parasitological cure in Chagas patients: a new serological multiplex immunoassay,” the authors evaluate anti-T. cruzi antibody dynamics to a set of antigens on the commercial MultiCruzi assay (InfYnity Biomarkers) in clinical trial specimens from the BENDITA trial to evaluate prediction of parasitological cure. There is a significant need for biomarkers that are more responsive to efficacious treatment considering the current outcomes used in drug trials either use PCR in peripheral blood, which is only positive in a fraction of chronic Chagas disease cases, or more recently a 20% decrease in signal from commercial ELISAs, which are not meant to be quantitative.

The MultiCruzi assay is unique in that multiple antigens printed on discrete dots within a microplate well capture antibodies from serum/plasma and generates a colorimetric reading of the individuals dots by image analysis. These colorimetric readings are normally applied to a cutoff, however, this paper employs an innovative method of serial dilutions to generate a sigmoidal curve (signal vs. titer), then generates a statistic of the halfway point along the linear portion of the dilutional curve (DF50) that can be compared between treatment timepoints from the trial samples (prior to treatment, 6 and 12 months). This is more efficient than end-point titers for each of the 15 antigens.

The findings of the study are positive, demonstrating significantly negative slopes of DF50 for certain antigens when plotted against time at 6 and 12 months in the treatment arm compared to the placebo arm. Furthermore, when compared to the previous method of 20% decrease in commercial ELISA signal, the signal change from MultiCruzi that the authors associate with efficacy of treatment was seen in a larger proportion of individuals in the treatment arms. The authors therefore conclude that the MultiCruzi assay is a better predictor of parasitological cure. These findings are consistent with previous publications of MultiCruzi in pediatric patients of the BENDITA trial, which are able to achieve seroreversion (to negative) (LJ Medina et al., 2021).

These results are promising and analytically demonstrate a robust method for evaluating dynamic antibody changes during treatment. The difficulty with this paper and discussion is that the authors seem to “jump the gate” when claiming this assay is a better predictor of parasitological cure. This is concerning when viewed through the lens that the first and corresponding authors are from the company that manufacturers this assay. As the introduction states, there are no tests for parasitological cure as PCR in peripheral blood only measures parasitemia (not potential intracellular reservoirs) and seroreversion (to negative) in adults only happens over years to decades, which is not feasibly demonstrated in the duration of this clinical trial. The outcome of parasitological cure in the adult specimen set referenced (13) are determined by PCR over 12 months. By this logic, this assay cannot infer predicted parasitological cure as this manuscript claims, only that this analytical method

of interpreting antibody decreases to specific antigens in the MultiCruzi assay is correlated with treatment and more responsive than conventional serology.

AUTHORS' ANSWER: We wish to thank Reviewer 1 for his positive analysis of our data. We can understand the "jump the gate" feeling and agree that we cannot yet make absolute conclusions regarding parasitological cure. We have therefore changed the wording i.e 'predicted cure' to 'response to treatment', 'recheck' to 'inconclusive' and 'no response' to 'no response to treatment'. We have also changed the title of the manuscript accordingly i.e to "Early antibodies decline assessment in Chagas patients following treatment using a new serological multiplex immunoassay".

This study is also lacking some key validation experiments, including defining precision of the DF50 in technical replicates high, medium, and low signal specimens to understand the variance that may occur within and between runs.

AUTHORS' ANSWER: For the sake of space, we had not included all experiments linked to the technical validation of the assay. The accuracy and reproducibility of the DF_{50} method depends on the number of serial dilutions. With only 3 dilutions (which was considered the minimal number with reduced cost and workload) we had to reduce the 4-parameter sigmoidal response curve to a 2-parameter response curve, by normalizing then fixing the Top to 100 and the Bottom to zero. Using 4 or more dilutions slightly increases accuracy and precision, as it allows to fit 'Top' and 'Bottom', next to the hillslope and DF_{50}.

Results of the Coefficient of Variation percentage analysis to assess intra-assay variability have been added as Supplementary Table 11 in the manuscript supplementary information. Analysis was performed following the testing in triplicates of 8 well-documented *Trypanosoma cruzi* positive samples having different reactivities (high, low and medium) with the antigens. Reproducibility data between runs were also generated as we have performed the testing of these samples in two different laboratories maximizing thereby the potential for variability in operational settings (interlaboratory settings: different laboratory, different operators (trained and untrained), independent runs, different reader, etc.). Correlation of the DF_{50} values for all antigens, all patients and at all timepoints (for 15 antigens, 201 patients and 3 timepoints, leading to 9045 calculated DF_{50} values), obtained in the two different laboratories show a high correlation between these results (correlation coefficient of 0.8965).

In addition, we calculated the Lin's Concordance Correlation Coefficient for each antigen of the MultiCruzi assay. Calculations showed a good correlation and agreement for each antigen.

Detailed data will be reported in a separate article dedicated to the technical validation of the dilution method used on two clinical trials samples analysis.

Lastly, this manuscript only documents the specific antigens used in the supplementary materials and does not discuss the findings in terms of host-pathogen interactions. This would be a fascinating biological study that deserves some mention, although I admit may be too broad for the focus of this paper.

AUTHORS' ANSWER: As stated by the Reviewer, we also believe that discussing the findings in terms of host-pathogen interactions could be very interesting but surely a too broad subject given the focus of the manuscript. This is surely an interesting point to follow, and the inclusion of data from additional trials will be helpful.

Overall, this study is innovative with exciting findings that are likely applicable to further clinical trials, but more in terms of analytical laboratory methods for assessing T. cruzi antibody dynamics versus actually demonstrating a method with adequate validation for assuring parasitological cure.

Minor Comments:
Line 52: This line makes it seem as if sudden death is the most likely outcome. For clarification, I would either remove this or add a few more words about the spectrum of diseases caused by T. cruzi infection.

AUTHORS' ANSWER: As the introduction has been shortnened, this part was removed.

Line 70: The use of "erratic" implies that the parasitemia has major fluctuations over time. It would be more appropriate to say "unreliable as a surrogate for the overall parasite load in other compartments and tissues."

AUTHORS' ANSWER: Thank you for this suggestion: we have changed this sentence as suggested.

Line 72: Saying PCR is inappropriate for assessing drug efficacy is a bold claim when it was used as an outcome measure in the BENDITA trial for which these specimens were generated. This is additionally concerning given the commercial interests of the authors. PCR is an analytically sensitive and specific method of direct parasite detection, the flaw is in the low parasitemia due to the biology of parasitic infection in humans. The authors note that the flaw in study design is that PCR detection in blood is not present in all people, but this does not make it inappropriate as for assessing drug efficacy for which PCR positivity is an inclusion criteria.

AUTHORS' ANSWER: We agree with the reviewer that, from a purely conceptual view, since PCR positivity is an inclusion criterion in the BENDITA study (and most others clinical trials as well), it is not necessarily inappropriate to use this technique to assess a treatment effect. We have therefore changed "inappropriate" to "has known limitations".

We would like to stress, however, that PCR limitations for treatment efficacy assessment in Chagas disease are now well recognized in the Chagas community. *T. cruzi* DNA as measured by PCR is indeed still used as an inclusion criterion in Chagas disease clinical trials despite the fact up to 80% of seropositive Chagas patients can be PCR negative depending on the regions. This is mainly due to the difficulty of assessing parasitological cure using conventional serology (i.e. seroreversion following treatment taking decades to occur in adults making conventional serology inadequate considering drug development timelines). However, the outcome measured using PCR in phase 2 proof of concept clinical trials is treatment failure (PCR positivity). Nevertheless, PCR is not an endpoint of treatment efficacy and is not accepted by regulatory authorities as a definitive measure of treatment success. While validated PCR is certainly the most sensitive method to directly detect parasite DNA in blood samples, the tropism of the parasite and the low parasitemia during the indeterminate chronic phase of Chagas disease in human do not make this technique satisfactory to give an idea of the overall parasite load in human body following treatment.

The search for new markers of parasitological cure and clinical benefit is part of DNDi overall Chagas disease strategy. Work on the MultiCruzi for Chagas disease is entirely sponsored by DNDi and is an integral part of this strategy to be able to deliver new treatment for these neglected patients.

Moreover, commercial interests in this specific field of neglected disease are certainly close to inexistant for this collaborative InfYnity Biomarkers / DNDi project. One should add that it is DNDi policy to ensure that new developments benefit the neglected patient in the first place.

Line 92: Should have a reference if including a statistic.

AUTHORS' ANSWER: Has been added accordingly

Line 108. These references are formatted differently for some reason.
AUTHORS' ANSWER: Thank you for the hint; these references have been reformatted.

Line 177: Figure 1a shows images of the antigens dotted on the well plate. In some high-intensity dots there is a trail or bleeding of color generated that drifts away from the dot and, in some instances, overlaps with other antigens or adds a tint to the whole well. Is this reviewed, corrected for or controlled in any way?

AUTHORS' ANSWER: Such operational issues are taken into account during the image analysis. Each microplate well is imaged and analyzed using the colorimetric sciREADER CL2 (SCIENION) reader. An integrated software identifies then calculates the pixel intensity for each spot. Therefore, the correction is made at two levels: per spot and per well as the background is also controlled.

Lines 179, 181: DF50 is formatted differently than previously in the paper.
AUTHORS' ANSWER: Thank you, this has been corrected to use the same format throughout the paper.

Reviewer #2 (Remarks to the Author):

The article represents a new confirmation of the usefulness of a new serological kit for the follow-up of patients with Chagas disease who have received etiological treatment.
The data obtained seem very interesting because, unlike current commercial kits, a more rapid drop in antibody levels is observed. It is well written, and the graphs show the results obtained in a very visual way. In any case, I think that before being so emphatic with the results, some extra analysis would be needed.
To evaluate such a test, the authors have used samples from patients included in the Bendita clinical trial, which evaluated different therapeutic schemes. Interestingly, all patients who have been exposed to the drug have had a response measured by the level of antibodies (more or less early depending on the treatment arm). Moreover, without a doubt this may represent a therapeutic effect of the medication.
In a second step, the authors reinterpret the data, and depending on the proportion of antigens that present a greater drop in this response, they venture to create a "cure" score.
What underlies healing is the presence or not of a viable parasite in the tissue, which on the one hand activates an immune response (from which healing can be inferred indirectly from its

quantification) as well as the inflammatory trigger that triggers tissue damage. and future morbidity and mortality. The methods used to date to measure therapeutic efficacy are the detection of DNA in peripheral blood using molecular biology (PCR). As the authors comment, this is done with the intention that if the treated patient has a positive post-treatment PCR, the patient is considered to have failed, since it is accepted that this DNA comes from an active infection.
Therefore, this method can provide an idea of what the kinetics of antibodies after treatment represent, but to attribute to them a "usefulness" as biomarkers of cure, it is necessary to associate these antibody levels with clinical/prognostic data (not feasible within of the time frame of any study) or the presence/absence of the parasite. The authors have an excellent opportunity that I believe they should not miss, which would be to analyze what the authors define as the risk of cure with the PCR results. This is essential since we can have the paradox of having patients with decreasing antibodies with sustained detection of parasite in the blood. In short, in order to understand this technique as a biomarker of cure, an analysis is needed that more or less indirectly confirms this outcome, and in this case, it would be the PCR.

AUTHORS' ANSWER: We thank the reviewer for this suggestion. We originally, on purpose, did not want to compare our results with those of the PCR for several reasons among which (non-exhaustive list):

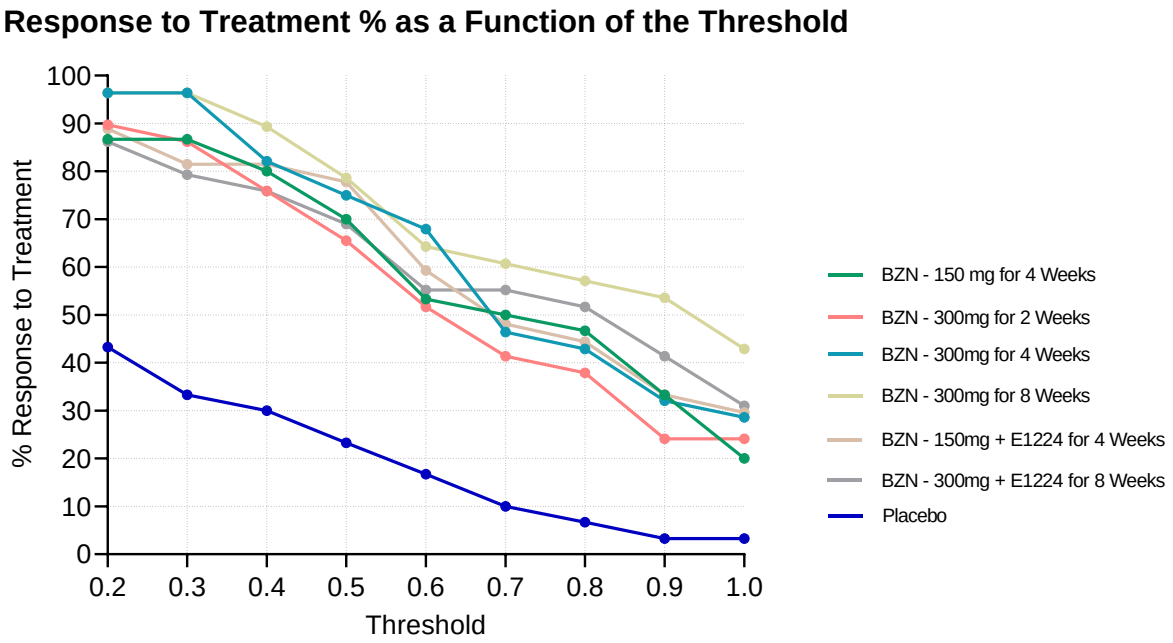
- PCR is used as an inclusion criterion to assess treatment failure after follow-up in most clinical trials for Chagas disease. However, PCR is not the gold standard surrogate marker for Chagas disease and is not accepted by health authorities. It is primarily used for practicality, given that seroreversion in adults takes decades, a timeline incompatible with drug development. The MultiCruzi assay described here is serology-based, the gold standard for Chagas disease monitoring, and aims to shorten the time needed to observe seroreversion in adults by identifying antibody decline, the first step towards seroreversion.
- It is widely accepted that a positive PCR is synonymous of *T. cruzi* DNA presence in blood and thereby, as stated by Reviewer 2, associated with an active *T. cruzi* infection, pending validated PCR method and adequate number of positive technical replicates; PCR can therefore be useful to assess treatment failure. The challenge is determining the meaning of a negative PCR as this is not a fundamental proof of absence of parasites in the body (*T. cruzi* is predominantly located in tissues and its presence in blood is "sporadic"). A series of negative PCR results do not necessarily confirm the absence of the parasite in the body tissues. Some longer follow-up studies have shown positive PCR after 4 or 5 years of follow-up while during the first years a sustained negative PCR was observed (see for example, Sulleiro E, et al. (2020) Usefulness of real-time PCR during follow up of patients treated with Benznidazole for chronic Chagas disease: Experience in two referral centers in Barcelona. PLoS Negl Trop Dis 14(2): e0008067. doi.org/10.1371/journal.pntd.0008067)
- PCR gives a binary outcome (presence or absence of *T. cruzi* DNA) that is not necessarily sustained in time e.g. that can be positive at baseline negative at 6 months follow-up and positive again at 12 months (see also point above). Moreover, the meaning of positive outcome at 6 months and not anymore at 12 months can be questioned. Parasiteamia is very low and most of the time hardly quantifiable (No samples quantifiable in the BENDITA trial at baseline, 6 and 12 months; All samples were lower than the LOQ of the method) and close to the LOD
- On the contrary, the MultiCruzi measures decline in antibodies already at 6 months, a decay that is sustained / further decreased at 12 months of follow-up. The presence of *T. cruzi* parasite will induce a sustained immune stimulation. Once the parasite is removed, the immune response will not be stimulated anymore, and antibodies will start to decay.

In short, both methods are measuring different outcomes: in our case, the MultiCruzi, is measuring antibodies decline as a surrogate of future seroreversion /treatment efficacy (seroreversion is accepted the regulatory authorities); PCR is looking at treatment failure measuring the presence of *T. cruzi* DNA in blood (sustained PCR negativity for one year following treatment) and is not accepted by the health authorities as a valid endpoint. It is therefore very difficult to make any formal comparison or incorporate PCR outcome data in the analysis of the MultiCruzi data; this will be only an arbitrary comparison.

Nevertheless, and notwithstanding the limitations described above, we have attempted to compare / assess any concordance between these two methods. The challenge is to define a cut-off that will fulfill both MultiCruzi and PCR. For the MultiCruzi assay, the cut-off refers to the number and intensity of reacting antigens needed to consider a treatment response. In the case of PCR, the cut-off is determined by the number of amplification cycles required to classify a result as positive or negative.

We first performed an analysis of the impact of the threshold used on the MultiCruzi outcome (added in the supplementary information as Supplementary Table 7 and Figure 4). Given the low number of patients in the placebo group as well as the lack of gold standard, a definitive threshold has not been fixed and the threshold for the current data was set at -0.3 based on maximizing the Youden index (S+Sp-1), which was calculated from ROC analysis. Notwithstanding this point, we noticed that whatever the threshold selected (between 0.2 and 1), the difference in percentage of response to treatment between treatment groups and the placebo group remains large. In all cases, a faster antibody decline in treated patients as compared to Placebo patients is observed; see Figure below. This figure has been added as Figure 5 in the manuscript.

Figure 5: Response to treatment as a function of the threshold used.



We have also added in the Supplementary Tables 9a and 9b with the comparison with PCR across the different thresholds.

In relation to the PCR results in the Placebo group, the best choice for the threshold is -0.7. This threshold corresponds to the following agreement with PCR (Figure 7):

Placebo: MultiCruzi Outcome (Threshold=-0.7)				
PCR	Response to Treatment	Inconclusive	No Response to Treatment	**PCR Total**
Sustained PCR < 0 during 12 months follow-up	3	2	24	**29**
PCR > 0	0	0	1	**1**
MultiCruzi Total	**3**	**2**	**25**	**30**

We added a formal comparison with the PCR results for all treatment groups. Similar to the analysis with the placebo group, while this can be done at different thresholds, we performed the comparison at the threshold of -0.7 (Figure 7, Supplementary Table 10).

All Treatment groups: MultiCruzi Outcome (Threshold=-0.7)				
PCR	Response to Treatment	Inconclusive	No Response to Treatment	**PCR Total**
Sustained PCR < 0 during 12 months follow-up	74	37	33	**144**
PCR > 0	15	9	33	**57**
MultiCruzi Total	**89**	**46**	**66**	**201**

It must be noted that the number of amplification cycles in PCR used to distinguish between positive and negative results is set arbitrarily and varies between regions (Duffy T et al. (2013) Analytical Performance of a Multiplex Real-Time PCR Assay Using TaqMan Probes for Quantification of Trypanosoma cruzi Satellite DNA in Blood Samples. PLoS Negl Trop Dis 7(1): e2000. doi:10.1371/journal.pntd.0002000). Consequently, adjusting this threshold can affect the number of positive or negative PCR results as well. In this context, comparing clinical trials that use PCR as an endpoint can be challenging.

PCR discordance: looking at the overall Chagas disease trials it is also important to notice the variability of outcome following treatment with benznidazole when measuring sustained parasite clearance for 1 year assessed using qPCR in different clinical trials. Compared to BENDITA trial results where 83% of the patients showed sustained clearance, in the MULTIBENZ trial (same treatment), 54% showed sustained clearance (Molina-Morant, D. et al. (2020) Efficacy and safety assessment of different dosage of benznidazol for the treatment of Chagas disease in chronic phase in adults (MULTIBENZ study): study protocol for a multicenter randomized Phase II superiority clinical trial. Trials 21: 328. https://doi.org/10.1186/s13063-020-4226-2). Obviously, this is also due to the variable parasite load (e.g. from very low close to the LOD in Bolivian patients, to high quantifiable in Brazilian patients) found in the blood of Chagas patients in different regions of Latin America.

We have added in the results section a paragraph highlighting this comparison and further discussed this issue (comparison MultiCruzi and PCR) in the discussion section as requested.

Reviewer #3 (Remarks to the Author):

Phase 2 trials of anti-parasitic interventions for chronic Chagas disease in adults are a major hurdle. Chagas research is hampered by the lack of understanding of disease pathogenesis and the factors associated with disease progression. A specific problem for phase 2 trials is quantifying parasitic cure. The parasite densities in blood in chronically infected adults are very low (usually less than 1 parasite per ml). These densities are only detectable by qPCR targeting the minicircle or satellite DNA. Drugs for Neglected Diseases initiative (DNDi) has done a series of phase 2 studies which have used “sustained parasite negativity” as the primary endpoint. The use of this endpoint is justified by the

ability to detect very low densities using PCR (PCR can in theory detect 1 parasite in 5 to 10ml of blood). Sustained parasite negativity is defined as undetectable parasites in blood from serial PCRs done over 1 year of follow-up. The problem with this endpoint is that not detecting parasites (a series of negative PCRs) does not mean absence of parasitaemia due to the very low densities. It is not possible to completely rule out the possibility of continued infection. There is a need for better endpoints to determine full cure, compare drugs, and optimise doses. In this paper, the authors explore the use of serological endpoints to determine cure.

MultiCruzi is a multiplexed antibody assay for T. cruzi containing 15 antigens printed onto 96-well plates. From serial dilutions, it is possible to derive a quantitative output for each of the 15 antigens. The authors have used samples from a phase 2 placebo-controlled randomised trial of different benznidazole regimens, some in combination with fosravuconazole (the BENDITA trial) to assess the predictive value of the MultiCruzi assay for "serological cure". The authors derive a 50% dilution factor (DF50: the dilution estimated to give 50% of the maximal signal intensity) for each antigen. This is estimated from 3 serial dilutions (1/50, 1/400, 1/3200). Changes in DF50 over time are compared between the randomised arms under a linear model. They conclude that samples taken 6 months after treatment initiation can predict "treatment efficacy", and this performance is much better than standard ELISA tests. They propose using the MultiCruzi output to define serological primary endpoints for "proof-of-concept" (phase 2) trials in chronic Chagas disease.

The data contained in this paper are very interesting and novel. Finding serological markers of cure that could be measured shortly after end of treatment would be a major advance for Chagas disease drug development. However, the statistical analyses presented are not appropriate and the conclusions are not warranted.

Major comments

1. In the abstract, the authors suggest including MultiCruzi in the primary endpoint for proof-of-concept trials (currently this is PCR). The authors do not carry out a formal comparison between Multicruzi and qPCR. This suggestion is not justified based on the presented analyses.

AUTHORS' ANSWER: As stated in our answer to Reviewer 2, we originally on purpose did not want to compare our results with those of the PCR since both endpoints are measuring different outcome i.e. antibody decline as a surrogate of treatment efficacy (seroreversion is the gold standard and is accepted the regulatory authorities) and treatment failure measured by PCR (sustained PCR negativity during one year following treatment) that is not accepted by the health authorities as a valid endpoint.

As the reviewer rightly stated, "… PCR is an imperfect endpoint …" and "the problem with this endpoint is that not detecting parasites (a series of negative PCRs) does not mean absence of parasitaemia due to the very low densities. It is not possible to completely rule out the possibility of continued infection". It is therefore very complicated to make any comparison or incorporate PCR outcome data in the analysis of the MultiCruzi data.

This being said and considering these limitations, we have made a first attempt to compare both MultiCruzi and PCR (See also our answer to Reviewer 2).

To be more appropriate, we have changed the sentence highlighting the potential of using the MultiCruzi as serology endpoint in future Phase 2 PoC clinical trials for Chagas disease. Only the

generation of additional new data will give information on the potential -or not- of this new methodology.

2. The primary endpoint in the original BENDITA trial was sustained PCR negativity. Although PCR is an imperfect endpoint, it has high specificity. Therefore, a positive PCR result during follow-up strongly indicates that the individual still has replicating parasites in their body. The analysis presented here does not incorporate any PCR outcome data. No reason is given for this omission. The serial PCR data were not used to calibrate the observed changes in antigen reactivity. By considering the PCR outcomes in the trial, it is clear that the derived thresholds for the antigen reactivities are mis-calibrated.

AUTHORS' ANSWER: Regarding the PCR data omission please see our also comment above.

We have to emphasize, however, that the thresholds for the antigen reactivities were determined based on ROC analysis for 'treated' vs 'placebo' groups. The maximum Youden index maximizes S + Sp – 1 and thus gives equal weight to S and Sp. However, implicitly this gives more weight to Sp as the number of patients in the Placebo group is only 30 compared to 180 in the combined treatment groups.

Taking into account that PCR results in the BENDITA study were all, except for a few exceptions, below the limit of quantification and Ct values varied between 27.71 and 39.9, we could not consider any statistical method for a reliable analysis. Unfortunately, the PCR could not be harmonized and repeated in the context of the present manuscript.

Nevertheless, we can vary the threshold which will agree with various (S, Sp) settings. We agree with the reviewer that this way of working is completely independent of the PCR results. To further clarify, we added in the supplementary section a comparison with the existing PCR results.

Similarly, and as depicted in our answer to Reviewer 2, the threshold for PCR (qPCR cut-off) can impact on the results as well (shift in Ct values). Moreover, the standard deviation at a high Ct (triplicate for each sample) can have also a major impact, increasing or decreasing the Ct by +3 or +4. One cannot rule out false positive or variation in the LOQ (LOD) either.

The proportion of patients predicted to be cured is too high. Table 1 shows the proportion of patients with "predicted cure" using the derived serological threshold from the analysis. In this table, it is estimated that 10 out of 30 patients given placebos were "cured" over one year follow-up.

But this contradicts the PCR data in the original trial publication in Lancet Infectious Diseases: only 1 individual out of 30 had sustained PCR negativity over 12 months. This implies that 9/10 of the patients with predicted serological cure had in fact positive PCR results during follow-up: i.e. they were not cured.

AUTHORS' ANSWER: We are not sure to be in a position to judge or determine if the proportion of patients predicted to be cured is too high or too low, whatever method used. We can only rely on the data and reports have already depicted rate of spontaneous cure up to 30% for example. Of course, if one compares with the PCR data this could be assumed to be too high. Nevertheless, considering the current amount of data available on the subject and the uncertainty linked to the various endpoints considered, we agree that one should be on the more conservative side. Therefore, and as stated to a comment of Reviewer 1, we have changed the term "cure in progress" to 'response to treatment' and

changed the title of our manuscript accordingly. As shown in the previous answer, the threshold can be changed to increase Specificity (to the cost of Sensitivity). When using the technical threshold as defined in the ROC analysis (-0.3), we indeed obtain an "unusual" spontaneous decline of antibodies in the placebo group, but this threshold remains to be further confirmed or modified by additional studies.

For example, at a threshold of -0.7, considering the Placebo group, only 3 patients are labeled 'response to treatment', 2 patients remain 'inconclusive' and 25 show 'no response to treatment'. This corresponds to a Sp = 25/30 = 83.3% and S = 3/30 = 10%. Concerning the PCR results in the Placebo group: 29 patients were PCR POS, of which 24/29 = 82.8% showed 'probably not responsive to treatment', while 3 patients with PCR POS were labeled 'response to treatment'. The patient with PCR NEG was labeled 'no response to treatment'; since a negative PCR cannot guarantee absence of parasites in the body, this could mean that PCR is overpredicting absence of treatment effect while the MultiCruzi shows absence of treatment effect. On the other hand, the reasons for PCR positivity in samples predicted to have an effect of treatment with the MultiCruzi remain to be determined and can have various reasons such as a threshold still to have optimized, outliers, as well as the difficulties to compare endpoints not measuring the same outcome. Overall, however, and despite the limitations described, it is fair to say that we found a good concordance between the PCR and MultiCruzi results (please see Tables above in our answer to Reviewer 2 on this topic).

3. There is value in comparing the longitudinal antibody patterns between the placebo and treatment groups. This is useful as a descriptive analysis. It is encouraging that there is more discriminative signal for the MultiCruzi assay compared to conventional serology. But (as mentioned in point 1) to show added value relative to PCR, and to argue that MultiCruzi should be used as a primary trial endpoint (instead of PCR), a lot more work needs to be done. There is already a huge and easily detectable difference in terms of PCR read-out between no treatment and relatively short treatment courses of benznidazole. The BENDITA study enrolled patients on the basis of a positive PCR at screening. 29 out of 30 patients randomised to placebo were consistently PCR positive throughout follow-up, whereas few of the benznidazole treated patients had positive PCRs during follow-up and those who did were generally positive at only a single timepoint (possibly indicating lower parasite densities much less that the lower limit of detection). Thus, PCR is very good at discriminating between treatment versus no treatment. The experience with fosravuconazole also indicates that PCR is good at determining whether a treatment does not work. What is less clear, however, is whether PCR can differentiate between slightly different treatments. One main output of the BENDITA trial was that 2 weeks of benznidazole appears to have similar efficacy as 8 weeks. If 2 weeks of treatment were sufficient, this would be a huge gain, as most side effects to benznidazole start in week 3, thus substantially improving adherence. I do not see how the analysis of the MultiCruzi data helps us to understand what the differences between these two treatment options are (if any).

AUTHORS' ANSWER: It should be emphasized that the BENDITA study did not have the aim to discriminate between treatments. As stated in the published BENDITA manuscript in Lancet Infectious Diseases, "patient numbers in this study were small, and the study was not powered for comparisons between the active-treatment groups. ..... The small number of patients in each treatment group precludes making firm conclusions regarding efficacy and safety data, necessitating further research in larger studies to confirm our findings." Accordingly, neither PCR nor the MultiCruzi data can help at that stage to discriminate between the treatment groups, but only between each treatment group and the placebo group.

We believe that an added value of the MultiCruzi is that it gives a dynamic and sustained response (continuous antibodies decline over time) between baseline 6 months and 12 months and is representative of the total body parasite burden. PCR gives a binary (on/off) outcome for parasite clearance in blood compartment only.

4. The presentation of the statistical analyses is inadequate. We thank the authors for sending their code and data. However, the code was written in SAS and was contained in a Word document. We note that SAS is proprietary software, and so we could not re-run the code. However, we did some exploratory analysis of the data in R. After looking at the data, we question some of the authors conclusions. Many of the DF50 estimates appear truncated, with values of either 0.1 (lowest value) or 6400 (highest value). We do not have the raw reactivity data but can guess that in these instances, the sigmoid fit "failed" (all observed reactivities were high, or all were low). For antigens 110 and 134, >90% of the values are truncated! We take this to mean that the dilutions used were not appropriate for these antigens. Antigen 99 has 60% of values truncated and antigens 36 and 101 have ~33% truncated. So out of 15 antigens, 5 provide little to no information (you cannot estimate change over time using a truncated datapoint). There are only 6/15 antigens which have less than 10% truncation. These issues are not reported in the paper. Nor does the statistical model take truncation into account. In summary the statistical analysis and presentation of the data is inappropriate.

AUTHORS' ANSWER: We do not think that the statistical analysis and presentation of the data is inappropriate or inadequate, on the contrary.

SAS is indeed a proprietary software widely used in the Pharma industry. It is approved by the FDA and was the software used in the original BENDITA study. We do not think this is the place to compare different data analysis softwares, and differences in the outcome are understandable given software differences and specificities.
This being said, we repeated the nested LMM analysis taking out all antigens that were not reactive at baseline i.e. with DF_{50} = 0.1 at baseline (and corresponding follow-up samples). The output was added to the supplement. Although the intercepts are much higher (which is obvious since we took out many 0.1 values) and the slopes changed to nearly the double, the overall conclusion still hold: all slopes (indicated by the Time(treatment) effects in the table below) are significantly different from zero with p-values < 0.0001. The slopes of the treatment groups are significantly different from the slope of the Placebo group, but no significant differences between treatment groups could be observed. Note also that 95%CIs can be calculated from the output below as effect size ± 1.96 x standard error. E.g. for BZN 150 4W, the slope is –0.2630 and the SE is 0.02030, resulting in a 95%CI of [-0.3028; -0.2232] (see Supplementary Tables 3 & 4, and Suppl. Figure 5). In the graph of the 95%CIs in one of the previous responses, the 95%CIs of the slopes of the nested LMM are shown visually.

Solution for Fixed Effects						
Effect	treatment	Estimate	Standard Error	DF	t Value	Pr > |t|
treatment	BZN _150_4	9.6292	0.2839	276	33.92	<.0001
treatment	BZN _300_2	9.7110	0.2871	282	33.83	<.0001
treatment	BZN _300_4	9.9073	0.3063	330	32.35	<.0001
treatment	BZN _300_8	10.0336	0.2913	282	34.44	<.0001
treatment	BZN+E1224_150+300_4	9.8218	0.3071	316	31.99	<.0001
treatment	BZN+E1224_300+300_8	10.0425	0.2971	308	33.80	<.0001
treatment	Placebo_0_0	10.1403	0.2836	273	35.76	<.0001
Time(treatment)	BZN _150_4	-0.2630	0.02030	3172	-12.95	<.0001
Time(treatment)	BZN _300_2	-0.2253	0.02057	3171	-10.95	<.0001
Time(treatment)	BZN _300_4	-0.2333	0.02216	3171	-10.52	<.0001
Time(treatment)	BZN _300_8	-0.2035	0.02088	3171	-9.75	<.0001
Time(treatment)	BZN+E1224_150+300_4	-0.2789	0.02216	3171	-12.59	<.0001
Time(treatment)	BZN+E1224_300+300_8	-0.2121	0.02139	3171	-9.91	<.0001
Time(treatment)	Placebo_0_0	-0.1007	0.02025	3178	-4.97	<.0001

For convenience in the calculation method, we verified any potential effect of truncation, and came to the conclusion that very low reactivities (DF_{50} = 0.1) and very high reactivities (DF_{50} = 6400) did not impede the overall conclusion.

To improve the fit with sigmoidal shape of the dilution curve, throughout the assay development, we have compared 8- vs. 3- dilutions and opted for 3 dilutions without significant change in accuracy.

Minor comments

1. Figure 1 shows well images and normalised signal intensities for two patients in the BENDITA trial at baseline, 6 months, and 12 months after treatment initiation. Are the actual data points being shown in panels b & c? It looks like they fit the sigmoid line perfectly, which makes me think they are not the actual data but predictions under the sigmoid model.

AUTHORS' ANSWER: They are actual experimental data points extended to 9 dilutions, fitted with the sigmoidal function and only serve to illustrate the shift in DF_{50} over time.

The DF50 line should be vertical instead of horizontal (the units of DF50 are dilutions, not the signal intensity).

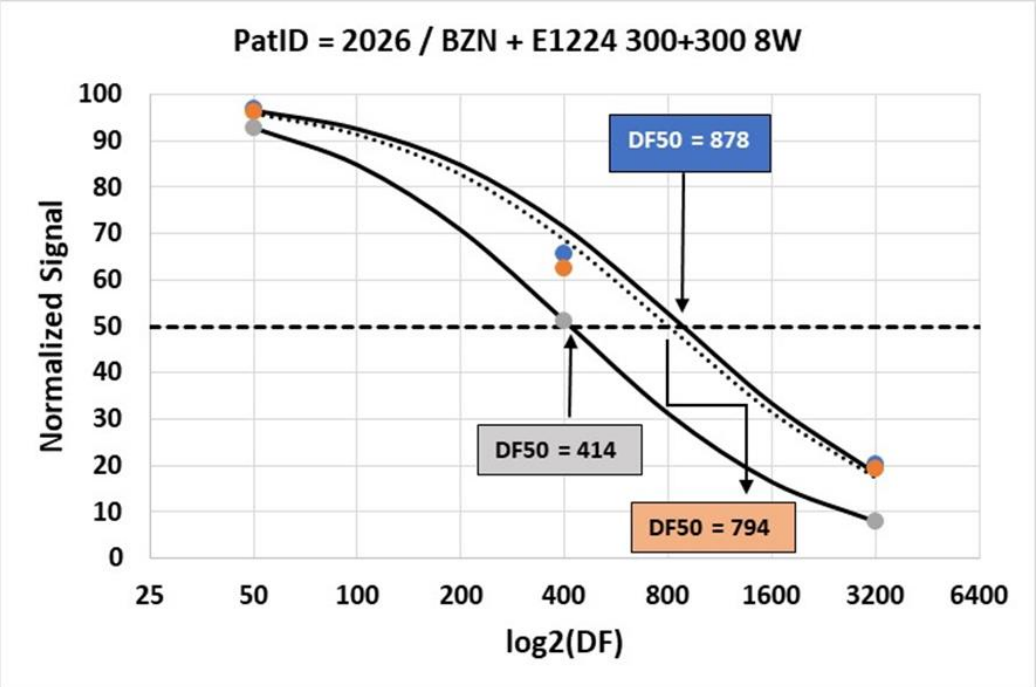
AUTHORS' ANSWER: We agree that the legend can be misleading. The horizontal line is drawn at 50% between Top (100) and Bottom (0) and the intersection with the sigmoidal curves allow to define the DF_{50}, which is indeed to be read from the horizontal axis. The graphs were modified accordingly.

The image in Figure 1 is useful and seems to suggest that the intensity measurement might be affected by "smearing" (not sure what the technical word for this is!). How is this captured in the digital data? More generally, how much noise is there (there are duplicate spots for each antibody which could help answer this question).

AUTHORS' ANSWER: This point was raised by Reviewer 1. These operational issues are taken into account during the image analysis. Each plate is imaged and analyzed using the colorimetric sciREADER CL2 (SCIENION) reader. An integrated software calculates the pixel intensity for each spot. In order to establish the net intensity for each antigen, we considered the mean value of the duplicated spots. Net signals correspond to the signal measured from each spot from which the background is subtracted.

2. As a general point, none of the Figures appear to show any raw data. Because no raw data are presented, it is very difficult to get a sense of the variability of the measurement, the goodness of fit and the true differences between the intervention arms. I think code and de-identified data should be made openly available for reproducibility. Code should be written in open source software (eg R or Python).

AUTHORS' ANSWER: We added an example of the raw data and the fitted results in the supplement (Supplementary Figure 3a and 3b), and another one is shown in the graph below (PatientID 2026, Antigen 14). To calculate the DF_{50}-values an Excel macro (programmed in Visual Basic for Applications) was used. The VBA macro can be made openly available. We believe most researchers have MS Office as standard software available.



3. On lines 68-72 the authors state: “However, T. cruzi is an intracellular parasite with cyclic and low parasitemia while it resides in the organs and during the indeterminate form of chronic Chagas disease. Its presence in the blood stream is thus erratic and does not represent the overall parasite load in other compartments and tissues. Therefore, the sensitivity and specificity of PCR methods are highly variable, making PCR inappropriate for assessing drug efficacy”.
qPCR is a well-established marker of drug failure. This is because PCR can detect very low numbers of circulating parasites in blood. It is not the specificity of PCR which is variable, it is the sensitivity. This sentence should be clarified and is not justified as PCR has not been compared with MultiCruzi.

AUTHORS’ ANSWER: As rightly stated by the Reviewer, PCR is used for assessment of treatment failure; PCR has, however, clear and recognized limitations to assess drug treatment efficacy, above all due a variable sensitivity between 40 and 70% in Chagas patients in the chronic phase of the disease. The MultiCruzi is dedicated to assess drug efficacy making it difficult to compare directly with PCR. We have further clarified the statement in the text and as mentioned by Reviewer 1 have changed “inappropriate” to “has known limitations”.

4. Most of the results are given in terms of significance testing. The emphasis on p-values is inappropriate. It is not very useful to know that the difference in reactivity changes over time between two groups is significant or not. This is primarily driven by the sample size. What we would like to know are the quantitative differences. For example, the statement “Slopes of all the treatment groups were significantly more negative than the slope of the placebo group ($p < 0.0001$), indicating that the antibodies decline more rapidly in treated patients than in patients administered the placebo.” The p-value is not useful here. No magnitude of effect is given. In addition most of the results are presented without confidence intervals and with superfluous precision in the estimates (for example a slope coefficient of -0.02766: do we need 5 decimals points?). I would suggest standardising the reactivity measurements so that the mean value at baseline for each antigen is 100. Then all results could be presented as a % change over time.

AUTHORS’ ANSWER: The magnitude of the effect is the magnitude of the slope (shown as the effect Time(treatment) in the LMM analysis). The p-values are indicative and exploratory in nature, being complementary to 95%CIs. In the Supplement we showed the SAS output tables, which give 5 decimal points, by default. Confidence intervals for DF_{50} dynamic are added in the supplementary. Being a retrospective study, our exploratory data were focused to evaluate differences only with the placebo group due to the limitations of sample size. Considering the low sample size in the BENDITA trial, it is worth mentioning that we still could find significant differences between treatment groups and placebo.

5. There seems to be arbitrary post hoc reporting for particular antigens. Why is antigen #11 special (main text) or antigen #3 (Figure 1)? There does not seem to be any formal comparison of the 15 antigens, are there any which look better than others?

AUTHORS’ ANSWER: The formal comparison between antigens is shown in Table 6 of the supplementary with the ROC analyses. The AUCs for each antigen are shown and can be compared among them. The relative importance of one antigen vis-à-vis the other depends on individual patients considering that each patient will develop a specific immune response.

6. Use of the Youden index implies that the same weight is being given to the sensitivity of a cut-off

value as to the specificity of the cut-off. But should that be the case here? Is it not more important to identify treatment failures?

AUTHORS' ANSWER: The maximum Youden index, defined as S + Sp –1, indeed gives the same weight to sensitivity as specificity. However, we should consider the sample size behind S and Sp, which is 180 and 30 resp, meaning that, implicitly more weight is given to the group with the lowest sample size, which is the Placebo group, and thus more weight is implicitly given to Specificity. Of course, another criterion than the maximum Youden index can be chosen for the antibody prediction algorithm. By increasing the threshold, we increase Specificity. We should keep in mind the very small Placebo sample size, and thus, by increasing the threshold we enhance the emphasis on the Specificity.

For a new treatment / drug, one wants to see treatment efficacy and not Treatment failure. That is also what regulatory authorities will consider. In that respect, PCR will never be considered as an endpoint especially in pivotal Phase 3 studies for Chagas disease, while the MultiCruzi that is a serology-based multiplex, pending additional confirmatory studies might.

7. Why not use a probabilistic model of cure? The set of rules proposed appear to be completely arbitrary and it is very unclear what the authors are trying to optimise for. As noted above, the predictions are known to be wrong (this predicts that 1/3 placebo patients are cured which contradicts known PCR data).

AUTHORS' ANSWER: We agree with the reviewer that the antibody prediction algorithm is at this stage still exploratory. However, we have based our rationale for the cut-off by performing ROC analysis. We also evaluated the use of a probabilistic model (logistic regression) but, to develop such a model, we need to define the outcome: cured or not cured, which is based on the definitions of the treatment groups, or based on PCR, both insufficiently reliable. Positive PCR is a surrogate for treatment failure and cannot be associated to cure.

8. The Introduction is very long, it could be shortened considerably.

AUTHORS' ANSWER: We agree that the introduction was very long. We have shortened it drastically as suggested.

9. Please clarify the objective of the study. The introduction section (Line 123) and the discussion section state different objectives.

AUTHORS' ANSWER: The objective of the study was clarified in the discussion to be aligned with the statement in the introduction.

10. Line 138 in discussion mentions a variability in the ELISA tests (used in the BENDITA trial) that was not mentioned in results or analysis. This should be clarified.

AUTHORS' ANSWER: The term variability is changed to inconsistency of ELISA results that was observed between two different techniques (Conventional and recombinant).

11. Line 111-112: the authors should state that there is actually a need for a reliable biomarker of early treatment response of any type. It does not necessarily have to be a serologic biomarker.

AUTHORS' ANSWER: Correct. It could also be a combination of markers and this has been changed accordingly in the text.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.

Nature Communications

Early assessment of antibodies decline in Chagas patients following treatment using a new serological multiplex immunoassay

Point-by-point rebuttal

Reviewer #1 (Remarks to the Author):

The comments are sufficiently addressed

AUTHORS' ANSWER: Thank you for your positive comments.

Reviewer #2 (Remarks to the Author):

My comments have been well argued and responded. Their publication may be very interesting for the scientific community.

AUTHORS' ANSWER: We wish to thank Reviewer 2 for his positive comments. We agree that the publication of these results may be of interest within the Chagas community and may play a major role in the development / registration of new drugs for this neglected disease.

Reviewer #3 (Remarks to the Author):

The authors have made some changes to the paper and have added an analysis using the PCR data (primary endpoint in the BENDITA trial).

I find some of the additional Figures very interesting and useful, especially Fig 5 which shows the effect of the threshold on the estimated "response to treatment".

AUTHORS' ANSWER: Thank you for your positive comments.

However, I remain of the opinion that this work is of insufficient quality for publication as currently presented.

AUTHORS' ANSWER: We are very surprised with this statement as we believe we have answered all comments thoroughly, made all changes and provided all data requested.

My main issues are:

• Data are not provided in an appropriate format resulting in work that is not reproducible. This is especially problematic because the main authors are from the company that make the assay (and therefore have a clear conflict of interest). Transparency in data is paramount.

AUTHORS' ANSWER: We provided all data, as requested by the reviewer, to the editors in a format that would enable him/her to repeat our analyses. The following documents were submitted via e-

mail (as we were not able to upload the VBA applications onto the highly secured server as proposed by Nature Comm's platform):

- The BENDITA raw data with the macro in VBA applied on the data in a Microsoft Excel document (.xlsm format)
- The VBA Codes for the DF_{50} calculations and Instructions for Use for the above document
- The DF50 detailed calculations method in a Microsoft Excel document (.xlsx format)
- The SAS Code for Linear Mixed Models in a word document

An application for Microsoft Excel (the macro in VBA) is available Free of Charge to anyone who uses Microsoft Office. Furthermore, SAS provides a free of charge basic software package that allows the reviewer to perform a linear mixed model analysis https://www.sas.com/en_be/software/on-demand-for-academics.html.

Finally, we thank Reviewer 3 for his critical comments on conflict of interest. As noted previously in our point-by-point rebuttal, commercial interests in the field of neglected diseases are close to non-existent. Our manuscript reports a collaborative project between InfYnity Biomarkers, Drugs for Neglected Diseases initiative (DNDi) and KULAK, a high-rank university with renowned statisticians and expertise in the field of in vitro diagnostics. We would like to stress that this work was sponsored by DNDi (two of the main authors are staff of DNDi) and thereby aligns with DNDi access policy which is to ensure that innovations and developments primarily benefit the neglected patients and thereby promotes the Open Access of data generated.

• The comparison between PCR and serology could be improved substantially. This is a classic "latent variable" problem. We can posit a true (unknown) "cured" versus "not cured" latent outcome for each individual in the trial. Both serology and PCR have different operating characteristics, notably +PCR has very high specificity for determining "not cured" and seroreversion has very high specificity for determining "cured". A latent variable model would allow for an objective estimation of the operating characteristics of both these tests conditional on the model assumptions. This would be a very interesting and highly publication worthy exercise. As currently written it seems like the authors are presenting highly ad hoc rules for determining "response to treatment" using DF50 units which are difficult to interpret. The section on PCR is really difficult to read.

AUTHORS' ANSWER: The comparison between PCR and serology was a common comment for all 3 reviewers and we believe that we have addressed it satisfactorily. Moreover, it should be reminded that this comparison is not the subject or aim of the manuscript either. All the data being available to the scientific community within this manuscript, any interested scientist could look further into the comparison of these endpoints if judged useful.

We have explained DF50 calculations in such a way that it is understandable by readers. In the first revision there were no comments on that topic. The interpretation algorithm is in a first part based on the threshold obtained by ROC analysis (maximized Youden index) for the pooled antibody reactivities (Figure 3), and in a second part on rules to integrate the reactive antibodies at baseline, to be able to make a medical decision on each individual subject. The linear mixed model clearly showed that there is a faster decline of antibodies in the treated patients compared to the Placebo group which is the first time that this is reported in adults because of the very slow decline in antibodies following treatment, but this result is 'on average'. Therefore, we established an interpretation algorithm that is in-line with the LMM results but can be used at the individual patient level. There is no doubt that such an interpretation algorithm for the individual patient level needs to be externally validated in other studies. But, as this is the first controlled study, there is always a moment needed for the development of such an algorithm.

To improve further our comparison between PCR and serology (MultiCruzi), we have added the following paragraph in the discussion section of the manuscript - "Our findings align with a recent analysis of qPCR data, which shows that Ct values remain stable over time in most placebo-treated patients, unlike those receiving other treatments in the BENDITA trial. Furthermore, all treatment

groups differ significantly from the placebo group, as demonstrated by the non-overlapping 95% Confidence Intervals for the estimated probability of cure in the Ct-based model for each treatment regimen compared to the placebo"- and the corresponding reference i.e. Watson, J. A. *et al.* Quantifying anti-trypanosomal treatment effects in chronic indeterminate Chagas disease: an individual patient data meta-analysis of two proof of concept trials (2024), that can be found in the Preprint server for health sciences MedRXiv at https://doi.org/10.1101/2024.07.14.24310398.

We thank the reviewer for his suggestion for using latent variables. Our biostatistics expert, Prof. Pottel, highlights the fact that LMM is a kind of latent variable model. See for example, Verbeke, G., & Molenberghs, G. (2017). Modeling through latent variables. *Annual Review of Statistics and Its Application*, *4*(1), 267-282. DOI: https://doi.org/10.1146/annurev-statistics-060116-054017.

• The authors completely ignore the key question around how many of the 15 antigens on the MultiCruzi are actually needed. As noted before, 5/15 antigens provide very little information (mostly truncated values). But how many of the remaining 10 are useful? Could all this just be done with 1 or 2 antigens?

AUTHORS' ANSWER: We indeed cannot tell with certainty at that stage what is the exact number of antigens needed to determine the response to treatment. However, we can say with a certain confidence that the 15 antigens currently in use in the multiplex assay provide needed information for 2 main reasons –diversity of antibody response at the individual level on the one hand and variability in antigens reactivity as a function of the geographical location /parasite DTUs on the other hand- that are further depicted below.

In the Supplementary Table 6 we show results from the ROC analysis, where 11 out of 15 antigens (as in fact 4/15, namely antigens 8, 9, 12 and 13 were not reactive with the tested samples) show an AUC which is significantly different from 0.5, showing diagnostic value, that indicates a discriminatory feature. So, we can clearly state that 11 out of 15 antigens are needed. It is important to say that each subject has a unique antibody 'profile' where not all of these 11 are necessarily reactive, but the 11 are needed because of the diversity of antibody profiles in this group of patients.

In the BENDITA study, 4/15 antigens are not reactive at baseline, and this is probably related to the geographic region, where specific strains are not appearing (but this may very well be the case in other regions). As this is a research assay, it can be further optimized by omitting non-reactive antigens, in case we discover that these antigens do not react in any other parts of endemic regions of Latin-America. However, this would be premature as we have shown for example that, while antigens#8 and #9 did not react at baseline in the BENDITA study (Bolivian population), they do react at baseline in children of two retrospective pediatric studies in Argentina (Medina et al., 2021). By sticking to one or 2 antigens, we would enhance the risk, depending on the population tested and its geographical localization, to be unable to detect any antibody decline. This is the purpose and advantage of using the Multiplex assay (looking at several antibodies separately and simultaneously) as compared to recombinant assays (for a single antigen) or conventional serology (looking at the global signal of a mixture of different antigens).

We have further elaborated on this topic in the part dedicated to the limitations of the study, in the discussion section of the manuscript.

• I question the statement “dilutions of 1/50, 1/400, and 1/3200 were found to be optimal for all fifteen antibodies assessed with the MultiCruzi assay”. 5 of the 15 provide very little information: ie dilutions are not optimal?

AUTHORS’ ANSWER: We are surprised with this statement as we have described in the paper that we extensively searched for the best compromise between accuracy and cost. Indeed, more dilutions, extending to over 6400, would benefit, only very marginally, the accuracy of the DF50-value, but to a greater cost and labor. In case all dilutions have a very high signal on a given antigen (thus even the 6400 dilution), this antigen is not responding to the treatment, so there is no need to further dilute to obtain a more accurate DF50-value. Also, as noted previously, to improve the fit with sigmoidal shape of the dilution curve, throughout the assay development, we have compared 8- vs. 3- dilutions and opted for 3 dilutions without significant change in accuracy. Moreover, additional information related to the process to derive an optimal dilution sequence can be found in the supplementary S1 file of Reference 34 that we quote specifically in our manuscript.

• Most of the paper is still framed in terms of significance testing which is not very useful or informative. The authors’ response to my previous comment on this seems to miss the point.

AUTHORS’ ANSWER: Clinical trials fail when the p-value for the main null hypothesis is > 0.05. So, statistical significance is of great importance to show that there is a difference of the rate of decline in antibodies between treated and placebo patients. As the BENDITA trial was not designed for this study, the power of the study was not high enough to show a difference between treated and placebo for the decline in a single antibody; but the strength of the current test and analysis is the multi-biomarker approach and the nested LMM that allows to show that the observed differences are also statistically significant.

Reviewer #3 (Remarks on code availability):
Not appropriate. They are using proprietary software and raw data are not made available. The code is given in the Word document. This is not of sufficiently high standard for reproducibility.

AUTHORS’ ANSWER: Please see our answer to first comment. Refer to change in manuscript. All codes were provided. We have added the sentence “free of charge” to make it clearer.

Reviewer #4 (Remarks to the Author):

I co-reviewed this manuscript with one of the reviewers who provided the listed reports. This is part of the Nature Communications initiative to facilitate training in peer review and to provide appropriate recognition for Early Career Researchers who co-review manuscripts.