PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Evaluation of seven recombinant VCA-IgA ELISA kits for the diagnosis of nasopharyngeal carcinoma in China: a case-control trial
AUTHORS	Gao, Rui; Wang, Lin; Liu, Qing; Zhang, Lifang; Ye, Yanfang; Xie, Shanghang; Du, Jinlin; Chen, Suihong; Guo, Jie; Yang, Mengjie; Lin, Chuyang; Cao, Sumei

VERSION 1 - REVIEW

REVIEWER	MINZHONG TANG WUZHOU RED CROSS HOSPITAL, WUZHOU, P.R.CHINA
REVIEW RETURNED	22-Jul-2016

GENERAL COMMENTS	Cao et al. present their case-control data evaluating seven recombinant VCA-IgA kits for the diagnosis of nasopharyngeal carcinoma in China. The authors conclude that three VCA-IgA kits had diagnostic effects equal to those of the standard kit and in combination with EBNA1-IgA can be used in future screening for NPC. Overall, this is a fairly well written paper with significant concerns with the results and discussion. There are a few grammatical or typing errors. Introduction 1. Page 4 line 22: “respectively.” It should have citation in this place. 2. Page 4 lines 28 to 37: The detail information for EBV capsid antigen should be in discussion part, instead of in introduction. It can be replaced by “Nowadays, several kinds of….” Methods 1. In page 8 lines 34 and 35 stated: “that p>0.05 was considered to be non-inferior”, but how cut point for the non-inferiority test based on the bootstrap approach were inferior to the standard kit are inadequately described. As the results show in Table 3 and Table 5, readers may misunderstand the means of p5, is that p>0.05 means non-inferior? Discussion 1. In page 11 line18: “(always 1)” what is the means for that? It should delete on the main text. 2. In page 11 line 54: “The NPC incident rate in the screening target population was relatively low”, is the “low” typing error? It should be “high”.

REVIEWER	Allan Hildesheim NCI, USA
REVIEW RETURNED	29-Jul-2016

GENERAL COMMENTS

Gao and colleagues report results from a study that attempted to formally compare the performance of various anti-EBV VCA IgA assays as screening tests to identify early stage NPC. The study is important and has the potential to inform decisions regarding assay choice for future NPC screening programs.
While the study addresses an important issue, there are some important limitations that should be considered:

1. EBV-based screening for NPC detection is needed for the detection of early (not late) stage NPC. Only 33 of the 200 NPC cases included in the present study had early stage disease. Thus, study power is very limited.

2. The comment above takes on added importance given that, in contrast to statements made by the authors throughout the paper, the sensitivity of some of the assays considered appears to vary considerably by disease stage. For example, for the BB assay (one of the 3 well-performing tests according to this report) has a sensitivity of 76% for early stage disease detection and 89% for late stage disease detection (Table 3). The lack of statistically significant differences between these 2 estimates is likely a reflection of the small sample size (low power) rather than a lack of difference between the estimates.

3. The authors do not report assay performance characteristics using the assay cutpoints defined by the manufacturers. These should be evaluated first, and represent a validation (or not) of the performance of these assays, as currently designed, for use in screening.

4. The optimized assay cutpoints evaluated by the authors are of interest, but represent a post-hoc evaluation that should be considered exploratory until findings are replicated using these new cutpoints in an independent study population. Thus, statements such as “Three recombinant VCA-IgA kits ….. can be used in future screening for NPC” (Abstract; last sentence) should be removed from the manuscript.

5. Table 3: It is unclear whether the cutoffs used for the 2 standard assays (EUROIMMUNE and EBNA1-IgA) were determined based on the 200 cases and 200 controls in this study (i.e., using the same approach to define “optimal” cutoffs to discriminate cases from controls for the 7 new assays evaluated in this study), or whether they were defined a-priori, using previous, independent experience by the authors. This is important to understand, since it has implications for how the results are interpreted.

6. Test-retest Reliability: The authors report ICCs only (Table 4). While ICCs are very imformative and important, it would also be of interest to include coefficients of variation (%CVs) and to evaluate the correlation between pairs of assays (using pearson or spearman correlation coefficients). Given that all of these assays target VCA, it would be informative to determine how well these assays correlated with each other.

Additional comments:

1. Study Population: Controls are all from Sihui but cases come from a broader catchment area. What proportion of the 200 NPC cases studied come from Sihui?

2. Table 1: It would be informative if the authors added a column to report the specific VCA antigen targeted by each of the 8 VCA assays considered.

3. Quality Control: The authors mention that 40% of samples were randomly selected for retesting as part of their QC. Were these replicate specimens tested in the same or different plates? Are the ICCs reported within or across plate ICCs? Additional details regarding their QC effort would be of interest.

4. Table 2: How do the authors explain the fact that the proportion of older NPC cases (defined as 50+) was higher for early stage (42%) compared to late stage (30%) cases?

5. Table 3: Please clarify that the column labelled "Average" contains the weighted (rather than simple average) of the sensitivities observed for early and late stage cases.

REVIEWER	Zilu Zhang, Biostatistician Havard Medical School and Havard Pilgrim Heath Care Institute, U.S.A
REVIEW RETURNED	07-Oct-2016

GENERAL COMMENTS

The research questions were well studied and the method was supported by some numeric results. However, this paper suffers from the problems below:

1. Abstract, Results: The authors indicate three "new" logistic regression models were built. It is unclear whether it is newer than the ones proposed in the reference paper "Establishment of VCA and EBNA1 IgA-based combination by enzyme-linked immunosorbent assay as preferred screening method for nasopharyngeal carcinoma: a two-stage design with a preliminary performance study and a mass screening in southern China" by Liu etc.; or the one proposed in this manuscript. I suggest to rephrase it by taking it out.

2. Strengths and limitations: control may suffer from the selection bias. Since hospital controls may have other characteristics or condition that led to hospitalization although they were convenient to select. Also, this section can appear after the discussion.

3. Introduction: 3rd line, the annual incidence rate of NPC in southern China is 25 per 100,000 person-years, not 25 per 100,000.

4. Methods, study population: Controls were selected from different hospital location and different time period (i.e. seasonal effect), it isn't clear whether these two factors will

affect the analyses.

5. Methods, statistical analysis: in paragraph 2, I assume VCA-IgA in the standard logistic formula is the standard kit EUROIMMUN. Although the authors mentioned it in the latter section, it will be clearer to state it in this section as well. Also, the authors did not explain why they did not include the baseline covariates in the model, although they may find complete frequency controls on age and gender etc.

6. Results, Table 2: some of the cell numbers is low (i.e. Age, NPC family history), Fisher's exact test is more appropriate in this situation. In the footnote, authors should state p<0.05 was considered as "statistically significant", should change the wording in the following context accordingly.

7. Result, page 7: it is unclear how different or non-different in sensitivities and specificities between the four kits and the standard kits. Is it by absolute value or statistical test?

8. Result, Table 3:
   a. Some of the kits were highlighted, but it is unclear why (i.e. EUROIMMU, EBNA1-IgA).
   b. Footnote 2 is not indicated in the table
   c. For footnote 2 and 3, the tests are performed on the same set of patients. The only information for comparing the sensitivities of the two diagnostic tests comes from those patients with a (+, -) or (- , +) result, Chi-square test is not appropriate. McNemar's test should be considered. Also, multiple comparisons should be taken into account, the significant level should be controlled, which is smaller than 0.05. BNV may no longer be statistically significant.
   d. Since the authors did not show p-values but confident interval, the footnote for the star "*" should be rephrased clearer.

9. Regarding the result of three logistic regressions, need significant level correction for the multiple comparisons.

10. Result, Table 5: the p-value should be replaces as "<0.001".

Typos (there may be more but I just list the following):
Page 8, the parenthesis is missing.
Page 11, the incidence rate is 50 per 100,000 person-years.

Overall, the paper carefully studied the proposed aims and methods both theoretically and numerically. However, some statistical test methods were problematic and need further explanation and substantial revision.

REVIEWER	Zhigang Haung Department of Epidemiology and Health Statistics, School of Public Health, Guangdong Medical University, China
REVIEW RETURNED	18-Oct-2016

GENERAL COMMENTS	This manuscript provides data evaluating the diagnostic effects of recombinant VCA-IgA ELISA kits for nasopharyngeal carcinoma in China， and it is interesting. Useful to discuss whether and how we can use these kits for NPC early detection and screening. However, important to acknowledge that the data presented in a high-exposure, high-incidence population and cutoff values from this study need be verified in prospective mass screening. Please think about the following questions: 1.Why choose these brands to evaluate the diagnostic effects of recombinant VCA-IgA ELISA kits for nasopharyngeal carcinoma? 2.The IgA antibodies against viral capsid antigen (VCA/IgA) and the IgA antibodies against EBV nuclear antigen 1 (EBNA1-IgA) were detected. The author need describe these two antibodies and tell the differences between them. 3.Is there any advantage to use Chinese recombinant VCA-IgA ELISA kits instead of the standard one?

REVIEWER	Dr.L.Jeyaseelan Christian Medical College Vellore, India - 632002
REVIEW RETURNED	07-Nov-2016

GENERAL COMMENTS	The authors have aimed at studying the 7 recombinant Elisa kits in the diagnosis of nasopharyngeal carcinoma. The gold standard test was EBNA1-IgA. Though the analyses were done mostly appropriately the following comments would be useful. 1. The authors have used AUC to compare between the 7 tests with the gold standard test EBNA1-IgA. The non inferiority margin used was 5%. They have used chi-square to test whether the difference is within 5%. However they have not provided the difference and 95% CI for the difference. This will provide us the information on what would be the maximum difference that would be obtained in the long run. May be that they should provide this using the boot-strap method. 2. The table 4 analyses, it is necessary, though it comes under the broader heading of methods to study the reliability. May be this information can be made as statements with the range of values. 3. The sub groups analyses of stage of cancer, especially Early stage (n=33) is a concern. The 95% CI for accuracy statistics are lower as compared to overall and advanced cancer. Though it is good that the authors have done this subgroup analyses, this needs to be discussed as a limitation of the tests as problem of numbers. Please provide the 95% CIs based on Bootstrap methods, for the Early stage accuracies. 4. The logistic regression analyses have shown a significant improvement in the AUC (that is 2 tests vs a single test). For example the AUC of KSB test increased from 0.945 (0.925, - 0.966) to 0.964 (0.947 – 0.981). The authors claim that the difference of

	about 1 or 2% is important?. I would argue that the single test is as good as two test combined together. We need to think about the cost issues and with the trade of in the gain in accuracy. This needs to be discussed in detail. I would not go by simple significance. 5. The authors have used to evaluate / compare the tests with the Gold standard (GS) test. Bland and Altman have provided warnings in using ICC in the reliability and validity studies. Please see the following REF. A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. J. M. Bland and D. G. Altman. Comput. Biol. Med. Vol. 20, pp. 337-340. 1990 Therefore, I would ask them to provide mean and difference plots for the combination of tests compared with the GS test. They need to provide bias, 95% CI and percentage error and other related statistics meant for reliability studies. Please refer the following paper. Bench-to-bedside review: The importance of the precision of the reference technique in method comparison studies – with specific reference to the measurement of cardiac output. Maurizio Cecconi1,2, Andrew Rhodes2, Jan Poloniecki3, Giorgio Della Rocca1 and R Michael Grounds2 Critical Care 2009, 13:201 (doi:10.1186/cc7129) Summary: Though the methods of analyses were focused to the objective, still the analyses need to provide appropriate analyses for reliability studies. Therefore, this paper may be accepted after extensive revision.

VERSION 1 – AUTHOR RESPONSE

Reviewer1:

1. Introduce: 1.Page 4 line 22: “respectively.” It should have citation in this place.
Re: Thanks for this suggestion. We have added the corresponding citation in this place. Please see the reference 16.

2. Introduce: 2. Page 4 lines 28 to 37: The detail information for EBV capsid antigen should be in discussion part, instead of in introduction. It can be replaced by “Nowadays, several kinds of….”
Re: As reviewer’s suggestion, we have moved the contents in Page 4 line 28 to 37 to the section of discussion. Instead, we introduced the several kinds of commercial VCA-IgA kits in this part.

3. Methods: In page 8 lines 34 and 35 stated: “that $p>0.05$ was considered to be non-inferior”, but how cut point for the non-inferiority test based on the bootstrap approach were inferior to the standard kit are inadequately described. As the results show in Table 3 and Table 5, readers may misunderstand the means of p5, is that $p>0.05$ means non-inferior?
Re: Thanks for this suggestion. $P<0.05$ was considered to be non-inferior for the non-inferiority test. In table 3 and table 5, we tried to use “*” to mark the kits which were inferior to the standard one, so we wrote “*p>0.05 was considered as statistically significant” in the footnotes. We had modified the expressions in these footnotes to prevent misunderstanding now.

In this study we use $\Delta=0.05$ as the cut point for Non-Inferiority Test for Paired ROC Curves. Here are the expiations (reference1-4 below),
Let θ1 and θ0 be the paired ROC curve areas for the new and the standard diagnostic kits and the Δ (θ0-θ1) be the pre-determined clinically meaningful equivalence limit. In this study, we let $\Delta=0.05$(cut point), because research always use $\Delta=0.05$ when Δ was hard to know. Then, we calculated the standardized differences λ1 and λ0 for the new and the standard diagnostic kits and let $\varepsilon= \Phi-1(\theta 0)-\Phi-1(\theta 0+\Delta)$.The hypothesis for non-inferior test is,
$H0: \lambda 1-\lambda 0<\varepsilon$ versus $H1: \lambda 1-\lambda 0\geq\varepsilon$, $\alpha = 0.\ 05$
If $p>0.05$, we should accept H0, which means the new kit was inferior to the standard diagnostic kit; if $p<0.05$, we should accept H1, which means the new kit was non-inferior to the standard one. We had emphasized the $\Delta=0.05$ in Method part, page 6 line 13.

1. Obuchowski N. Testing for equivalence of diagnostic tests. Am JRadiol 1997;168:13–7.
2. Liu JP, Ma MCh, Wu ChY, et al. Tests of equivalence and noninferiority for diagnostic accuracy based on the paired areas under ROC curve. Statist Med 2006;25:1219–38.
3. Zhou XH, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: Wiley, 2002:188–92.
4. Chen WZ, Zhang JY. Application of non-inferiority test for diagnostic accuracy under the areas of ROC based on bootstrap approach and its macro programming development. Modern Prev Med (Chinese) 2010;37:3009–10.

4. Discussion: 1. In page 11 line18: “(always 1)” what is the means for that? It should delete on the main text.
Re: Thanks for this question. The meaning of this sentence is that all of the original cutoff values were set to 1 according to the kits instructions. We have deleted the “(always 1)” and amended the description by the suggestion of the reviewer. Please see it in line18th, page 11.

5. Discussion: 2. In page 11 line 54: “The NPC incident rate in the screening target population was relatively low”, is the “low” typing error? It should be “high”.
Re: Thanks for this question. Although NPC incidence (about 50/100, 000) in the screening target population is higher than that in other population. Compared with other common diseases, such as hypertension and diabetes（the incidences more than 1%）， the NPC incidence in the screening target population is still relatively low. I have added some explanation in 54th line, page 11 to make it easier to be understood.

Reviewer2:

1. EBV-based screening for NPC detection is needed for the detection of early (not late) stage NPC. Only 33 of the 200 NPC cases included in the present study had early stage disease. Thus, study power is very limited.

Re: Thanks for this question. EBV-based screening for NPC detection is indeed aimed for detection of early stage patients. In theory, recruitment of more early stage NPC cases is better for evaluation of the NPC early detection values of EBV antibodies. Due to the low percentage (less than 20%) of early stage in clinic, recruitment of early stage NPC cases is more difficulty and we can only collect 33 early stage NPC participants in our study period in Sun Yat-sen University Cancer Center (SYSUCC).

The articles about the relationship between EBV-related antibody titers and the NPC clinic stages are limited. However, we found that the EBV-related antibodies including VCA-IgA have equal diagnostic ability for early and advanced stage NPCs (reference1-2 below), and the elevated EBV antibodies even appeared several years of NPC occurrence (reference3 below). The data in our study also verified this point. There was no significant difference of EBV antibody level between early and late NPC patients in this study. The difference of rRod of VCA-IgA and EBNA1 for early and advanced stage NPCs were compared by Mann-Whitney U Test. The P values were 0.255 and 0.101, respectively. So, this could prove VCA-IgA in this study have the same sensitivities for detection of early and advanced stage NPC patients although only 33 early stage cases were included in this study. We have added the comments in the Discussion, Page 12, line 14th.
1. Stolzenberg MC, et al. Purified recombinant EBV desoxyribonuclease in serological diagnosis of nasopharyngeal carcinoma. Int J Cancer. 1996 May 3;66(3):337-41
2. Zeng Y, Zhong JM, Li LY, et al. Follow-up studies on Epstein-Barr virus IgA/VCA antibody-positive persons in Zangwu County, China. Intervirology 1983;20:190–4.
3. Cao SM, Liu Z, Jia WH, et al. Fluctuations of Epstein-Barr virus serological antibodies and risk for nasopharyngeal carcinoma: a prospective screening study with a 20-year follow-up. PLoS One 2011;6:e19100

2. The comment above takes on added importance given that, in contrast to statements made by the authors throughout the paper, the sensitivity of some of the assays considered appears to vary considerably by disease stage. For example, for the BB assay (one of the 3 well-performing tests according to this report) has a sensitivity of 76% for early stage disease detection and 89% for late stage disease detection (Table 3). The lack of statistically significant differences between these 2 estimates is likely a reflection of the small sample size (low power) rather than a lack of difference between the estimates.
Re: Thanks for this good suggestion. We checked the result of three well-performing kits in table 3.The P value was 0.065 for BB (screenshot from SPSS below), 0.321 for HA and 0.057 for KSB. The sensitivity of KSB for early stage NPC was higher than that of advanced stage in absolute value. The P values of HA and BB were not near the threshold (0.05). Though the lack of early stage NPC patients cannot be denied, we think the purpose of grouping was merely to make our study stricter. Only when the difference between two stages was statistically significant in our research, the kit will be ignored. We have added the comments in the Discussion, Page 12, line 14th.

3. The authors do not report assay performance characteristics using the assay cutpoints defined by the manufacturers. These should be evaluated first, and represent a validation (or not) of the performance of these assays, as currently designed, for use in screening.
Re: Thanks for this suggestion. Actually, we have evaluated the cutoffs in the manufacturers first (table below). But as explanations in the discussion part said, we found the sensitivities and specificities were not reasonable. Since AUC was the key target in our study, we decided to put the table with AUCs and new cutoffs-defined with the largest Youden Indices chosen from ROC-in the result part.
Kits Sensitivity (%)(95%CI) Specificity (%)(95% CI)
VCA-IgA

BB	83.0 (77.8-88.2)	94.5 (91.3-97.7)
BNV	84.5 (79.5-89.5)	94.0 (90.7-97.3)
GBI	71.5 (65.2-77.8)	93.0 (89.5-96.5)
HA	86.5 (81.8-91.2)	87.0 (82.3-91.7)
HK	85.0 (80.1-89.9)	85.5 (80.6-90.4)
KSB	78.0 (72.3-83.7)	92.5 (88.8-96.2)
ZS	59.0 (52.2-65.8)	95.5 (92.6-98.4)
Euroimmun	91.5 (87.6-95.4)	79.5 (73.9-85.1)
EBNA1-IgA	89.5 (84.7-93.3)	90.0 (85.8-94.2)

4. The optimized assay cut points evaluated by the authors are of interest, but represent a post-hoc evaluation that should be considered exploratory until findings are replicated using these new cut points in an independent study population. Thus, statements such as “Three recombinant VCA-IgA kits…..can be used in future screening for NPC” (Abstract; last sentence) should be removed from the manuscript.
Re: Thanks for this good suggestion. The main objective of this study was to evaluate the diagnostic effects of these recombinant VCA-IgA kits and to provide more choices for NPC screening in the future. As you advised, the cutoffs cannot be defined in our study, so we modified the expressions in lines 27th and 39th in page 3 and lines 34th in page 10. News cutoffs were provided but researchers need to replicate them in more studies.

5. Table 3: It is unclear whether the cutoffs used for the 2 standard assays (EUROIMMUNE and EBNA1-IgA) were determined based on the 200 cases and 200 controls in this study (i.e., using the same approach to define “optimal” cutoffs to discriminate cases from controls for the 7 new assays evaluated in this study), or whether they were defined a-priori, using previous, independent experience by the authors. This is important to understand, since it has implications for how the results are interpreted.
Re: Thank you for this question. The cutoffs in table 3 used for the 2 standard assays were also determined based on the 200 cases and 200 controls in this study. These cutoffs were defined with the largest Youden Indices. The formula for the 2 standard assays used in this study was priori established and already used in an NPC cohort in Guangdong, China now. We used this formula to define low/medium/high-risk people in this NPC cohort (mass screening). But we didn’t established a-priori cutoff for distinguishing between NPCs and controls of this formula. So, in table 5, cutoff for the standard formula was defined based on the 200 cases and 200 controls too.

6. Test-retest Reliability: The authors report ICCs only (Table 4). While ICCs are very informative and important, it would also be of interest to include coefficients of variation (%CVs) and to evaluate the correlation between pairs of assays (using pearson or spearman correlation coefficients). Given that all of these assays target VCA, it would be informative to determine how well these assays correlated with each other.
Re: Thanks for the suggestion. We found the distributions of antibodies didn’t conform to the normal distributions and they didn’t belong to the qualitative variables either. The coefficients of variation, Pearson Correlation coefficient and the Kappa analysis might not be suitable. So we decided to use the ICC at last. Any index has its limitation. We calculated spearman correlation coefficients indices to evaluate the test–retest reliability of these kits as follow, but we felt puzzled about the “%CVs”. Did it mean the CV of 400 samples of each kit; or CVs for 200 case and 200 controls of each kit? The distribution of rOD was skewed. CV might not be right for the data.
The Spearman Correlation Coefficients of Test–retest reliabilities of
Eight VCA-IgA kits and the EBNA1-IgA kit

Kits	SCC	p*
VCA-IgA		
BB	0.92	<0.001
BNV	0.87	<0.001
GBI	0.85	<0.001
HA	0.87	<0.001
HK	0.86	<0.001
KSB	0.71	<0.001
ZS	0.66	<0.001
EUROIMMUN	0.82	<0.001
EBNA1-IgA	0.93	<0.001

* $p<0.05$ was considered as statistically significant correlation.

Additional comments:

1. Study Population: Controls are all from Sihui but cases come from a broader catchment area. What proportion of the 200 NPC cases studied come from Sihui?

Re: Thanks for this question. Though cases were collected in Guangzhou and controls were collected in Sihui, both of them were from the core parts of high-endemic area of NPC (black parts shown in the map below, reference 1 below). We checked the proportion of cases and found that 70% of them were from Guangzhou and nearby cities, including Sihui. Furthermore, we did not found evidences that there were differences in the morbidity of NPC, the injection of EBV or the titers of EBV antibodies among these places (reference 1-4 below). We have added the comments in the Discussion, Page 12, line 14th.

1. Jia W-H, Huang Q-H, Liao J, et al. Trends in incidence and mortality of nasopharyngeal carcinoma over a 20–25 year period (1978/1983–2002) in Sihui and Cangwu counties in southern China. BMC Cancer 2006;6:178–85.
2. Li K1, et al. Time trends of nasopharyngeal carcinoma in urban Guangzhou over a 12-year period (2000-2011): declines in both incidence and mortality. Asian Pac J Cancer Prev. 2014;15(22):9899-903.
3. Zhang LF, et al. Incidence trend of nasopharyngeal carcinoma from 1987 to 2011 in Sihui County, Guangdong Province, South China: an age-period-cohort analysis. Chin J Cancer. 2015 May 14;34(8):350-7
4. Xiong G, et al. Epstein-Barr virus (EBV) infection in Chinese children: a retrospective study of age-specific prevalence. PLoS One. 2014 Jun 10;9(6):e99857.

2. Table 1: It would be informative if the authors added a column to report the specific VCA antigen targeted by each of the 8 VCA assays considered.

Re: Thank you for the suggestion. We had tried our best to contact with the companies of seven recombinant VCA assays but companies only provide limited information, maybe because of patents or trademarks. The capsid proteins in the EUROIMMUN kit were extracted from the pyrolysis products of human B lymphocytes (P3HR1 cell line) infected by EBV which also was a combined native capsid protein of EBV. So, unfortunately, we cannot get the detail VCA antigens of these assays.

3. Quality Control: The authors mention that 40% of samples were randomly selected for retesting as part of their QC. Were these replicate specimens tested in the same or different plates? Are the ICCs reported within or across plate ICCs? Additional details regarding their QC effort would be of interest.

Re: Thanks for this question. 10% of the samples were randomly selected for retesting. We designed an instruction for testing replicate specimens. Since we got 5 plates for each brand, we left 8 holes in each plate and used half of these holes for testing replicate specimens in same plate and the other four for specimens from different plates. When all samples were tested, the data from 400 samples and 40 retesting samples were collected and then the ICC was calculated. It was not within or across plate ICC but a kind of “mixture” ICC. All plates in one brand were from a same batch and we had followed the instruction to prevent potential confounding factors.

4. Table 2: How do the authors explain the fact that the proportion of older NPC cases (defined as 50+) was higher for early stage (42%) compared to late stage (30%) cases?
Re: Thanks for this question. Serum specimens were continuously collected from hospitalised patients with NPC in the Sun Yat-sen University Cancer Center (SYSUCC) from January 2013 to June 2013. When 200 qualified cases (conformed to inclusion criteria) were collected, the age distributions of early stage and advantage stage were clear and definite. There was no human choice in this process. What’s more, no significant difference in age was found in the two groups by chi-squared tests in table 2.

5. Table 3: Please clarify that the column labelled “Average” contains the weighted (rather than simple average) of the sensitivities observed for early and late stage cases.
Re: Thanks for this question. The sensitivities in this column was not calculated by an average of two parts, but were calculated by all cases together. So we think the word “total” is more suitable than the word “average” here. Sorry for the confusion and we have modified these words in table 3 and 4.

Reviewer 3:

1. Abstract, Results: The authors indicate three “new” logistic regression models were built. It is unclear whether it is newer than the ones proposed in the reference paper“Establishment of VCA and EBNA1 IgA-based combination by enzyme-linked immunosorbent assay as preferred screening method for nasopharyngeal carcinoma: a two-stage design with a preliminary performance study and a mass screening in southern China” by Liu etc.; or the one proposed in this manuscript. I suggest to rephrase it by taking it out.
Re: Thank you for the suggestion. We used “new” to explain these three models were derived from three new used VCA-IgA kits. But just as reviewer said, all these models were based on logistic regression model. So we deleted the “new” in line 22nd, line 29th, line 36th and 44th in page 3; line 31st in page 10.

2. Strengths and limitations: control may suffer from the selection bias. Since hospital controls may have other characteristics or condition that led to hospitalization although they were convenient to select. Also, this section can appear after the discussion.
Re: Thank you for this good suggestion. This study was a Single-center study and controls were from hospital. Though we had chosen physical examination people as controls, it might still have indiscoverable bias as the reviewer said. We analyzed the feasibility of finding controls from community in Guangzhou, China at first but found it harder to accomplish. We modified the limitations tips and added the comments in the Discussion, Page 12, line 14th.

3. Introduction: 3rd line, the annual inciden ce rate of NPC in southern China is 25 per100,000 person-years, not 25 per 100,000.
Re: Thank you for the suggestion. We have modified the presentations in the Introduction and Discussion.

4. Methods, study population: Controls were selected from different hospital location and different time period (i.e. seasonal effect), it isn't clear whether these two factors will affect the analyses.
Re: Thanks for the suggestion. There was no strong evidence showing that the virus titres in the host can be affected by weather, temperature or season. Both the two hospitals were located in core parts of high-endemic area of NPC (black parts shown in the map below, reference 1 below). We did not found evidences that there were differences in the morbidity of NPC, the injection of EBV or the titers of EBV antibodies among these places (reference 1-4 below). We have added the comments in the Discussion, Page 12, line 14th. Furthermore, as commercial kits, the testing results should not be affected by external factors outside the specs. So we don't think the two factors will affect the analyses.

1. Jia W-H, Huang Q-H, Liao J, et al. Trends in incidence and mortality of nasopharyngeal carcinoma over a 20–25 year period (1978/1983–2002) in Sihui and Cangwu counties in southern China. BMC Cancer 2006;6:178–85.
2. Li K1, et al. Time trends of nasopharyngeal carcinoma in urban Guangzhou over a 12-year period (2000-2011): declines in both incidence and mortality. Asian Pac J Cancer Prev. 2014;15(22):9899-903.
3. Zhang LF, et al. Incidence trend of nasopharyngeal carcinoma from 1987 to 2011 in Sihui County, Guangdong Province, South China: an age-period-cohort analysis. Chin J Cancer. 2015 May 14;34(8):350-7
4. Xiong G, et al. Epstein-Barr virus (EBV) infection in Chinese children: a retrospective study of age-specific prevalence. PLoS One. 2014 Jun 10;9(6):e99857.

5. Methods, statistical analysis: in paragraph 2, I assume VCA-IgA in the standard logistic formula is the standard kit EUROIMMUN. Although the authors mentioned it in the latter section, it will be clearer to state it in this section as well.
Also, the authors did not explain why they did not include the baseline covariates in the model, although they may find complete frequency controls on age and gender etc.
Re: Thank you for the advice. I have added brackets and instructions behind "VCA-IgA" in the formula in 26th line, page 6. Hope it can reduce misunderstanding. Instead of an etiology study of epidemiology, our research is a diagnostic test. The purpose of this study was not establish risk model for NPC but provided appropriate combinations of different VCA-IgA and EBNA1-IgA kits to get good diagnostic effects for clinic and mass screening use in the future. So including epidemiologic features might not be suitable. The complete frequencies between cases and controls on age, gender and NPC history were just used to show that these people had comparability.

6. Results, Table 2: some of the cell numbers is low (i.e. Age, NPC family history), Fisher's exact test is more appropriate in this situation.
In the footnote, authors should state p<0.05 was considered as "statistically significant", should change the wording in the following context accordingly.
Differences in early- and advanced-stage NPC were compared by Chi-Squared tests (N=200). If the cell number was too low (the number in more than 1/5 cells had expected count less than 5 or any cell cells had expected count less than 1) when using this test, we would use Fisher's exact test instead of the Person Chi-Square test.
Re: Thank you so much for these important tips. As the reviewer said, we found he "Age", "Smoking" and "Drinking" were suited to Person Chi-Square Test (0 cell have expected count less than 5) and the "NPC history" was suited to Fisher's Exact Test (1 cell have expected count less than 5. The minimum expect count is 2.64). Since I thought Chi-Squared test (X2-Test) contains Person Chi-Square Test, Continuity Correction Test, Likelihood Ration Test, Fisher's Exact Test, and McNemar

Test, I used the Chi-Squared test as a general name. We had modified these presentations in “Method” part and “Results” part. Now the items which used Fisher’s Exact Test or other tests have been marked.
We also have changed “significant” into “statistically significant” in footnotes of table 2-5.

7. Result, page 7: it is unclear how different or non-different in sensitivities and specificities between the four kits and the standard kits. Is it by absolute value or statistical test?
Re: Thanks for the good question. At first the differences in the sensitivities and specificities of EUROIMMUN and other kits were compared by McNemar’s test, but now we think these comparisons were not necessary. The AUC for each kit is fixed and it’s the golden target for evaluating diagnostic efficacy. But the sensitivity and the specificity are correlative and can be changed by cutoffs. It is meaningless for comparison of them alone. Only when one of the sensitivity or the specificity is fixed, we can compare the other one. So we deleted the comparisons between the sensitivities and specificities of EUROIMMUN and other kits

8. Result, Table 3:
a. Some of the kits were highlighted, but it is unclear why (i.e. EUROIMMU, EBNA1-IgA).
b. Footnote 2 is not indicated in the table
c. For footnote 2 and 3, the tests are performed on the same set of patients. The only information for comparing the sensitivities of the two diagnostic tests comes from those patients with a (+, -) or (- , +) result, Chi-square test is not appropriate. McNemar’s test should be considered. Also, multiple comparisons should be taken into account; the significant level should be controlled, which is smaller than0.05. BNV may no longer be statistically significant.
d. Since the authors did not show p-values but confident interval, the footnote for the star “*” should be rephrased clearer.
Re: Thanks for these tips.
a. Among the seven new kits, three kits (GBI, HK, and ZS) were written in italics to show their AUCs were not as high as the standard kit while the other four (BB, BNC, HA and KSB) were written in non-italics. Among the four kits, the BNV had differences between early- and advanced- stage NPC in the sensitivities, so it was written in non-bold while the other three were written in bold. These three kits, the standard EUROIMMUN kit and the EBNA1-IgA kit which would be used in the combination step were written in bold.
As the reviewer said, the highlighted words might cause misconception and we have deleted them now.
b. Thank you so much and we have modified.
c. For question 7, we deleted the footnote 3 now. For footnote 2, the test was between early stage and advanced stage in one same kit (if p<0.05, “*” shown in “Advanced stage” column). After confirming the cutoffs, we calculated the sensitivities of early stage and advanced stage for each kit. The Person’s Chi-square test was suit to the comparison (0 cell have expected count less than 5). Just as the mistake I had made in table 2, I treated the Chi-Squared test (X2-Test) as a general name for Person’s Chi-square test and McNemar test. We have rephrased them now. “BNV” had different sensitivities in two stages. The significant level for two stages didn’t need to be adjusted because they were only compared once for each kit.
d. We have marked all “*” in red and modified some explanations in footnotes.

9. Regarding the result of three logistic regressions, need significant level correction for the multiple comparisons.
Re: Thanks for the suggestion. The diagnostic efficacy of each formula was evaluated by AUC. We sorted the AUCs of the three regressions and compare the smallest one (HA+EBNA1-IgA) to the

standard one. We found there was no difference between the AUCs of these two models. The AUCs of the other two new formulas were bigger than HA, so we can get the conclusion that the difference of the AUCs between these two models and the standard one were not statistically significant without comparison. We only compared them for one time, the significant level don’t need be adjusted. We calculated p values of AUCs between BB, KSB and the standard model in table 5. If the significant level for compares of AUCs between each new combination and the standard one be changed to 0.017 (α=0.05/3), the results would have no change.

10. Result, Table 5: the p-value should be replaces as “<0.001”.
Re: Thank you so much and it has been modified.

11. Typos:
Re: Thanks and we have tried our best to find out and modify these typos.

Reviewer 4:

1. Why choose these brands to evaluate the diagnostic effects of recombinant VCA-IgA ELISA kits for nasopharyngeal carcinoma?
Re: Thanks for this question. When we decided to do this research, we use the website of China Food and Drug Administration (http://eng.sfda.gov.cn/WS03/CL0755/) to find all registered legal VCA-IgA ELISA kits in China. We found eight brands of VCA-IgA ELISA kits in June, 2013 and all of them were recombinant kits. But the specification of one of these kits (48-well plates) was different with the other Chinese kits and the standard ones (96-well plates). In order to keep the accuracy and comparability of this research, we use these seven kits to evaluate the diagnostic effects.

2. The IgA antibodies against viral capsid antigen (VCA/IgA) and the IgA antibodies against EBV nuclear antigen 1 (EBNA1-IgA) were detected. The author need describe these two antibodies and tell the differences between them.
Re: Thanks for this suggestion. This study and our previous researches have shown that the combination of VCA/IgA and EBNA1-IgA could increase NPC diagnostic accuracy. VCA-IgA and EBNA1-IgA are two antibodies corresponding to EBV lytic-cycle proteins and latency gene products, respectively. The detailed descriptions were in the third paragraph of Introduction part and the third paragraph of Discussion part.

3. Is there any advantage to use Chinese recombinant VCA-IgA ELISA kits instead of the standard one?
Re: Thanks for this question. In this study, three Chinese recombinant kits can be substituted for the standard kit, and their combinations can be used in the early detection of and screening for NPC. The realistic significances for this study is that the prices of these recombinant kits were only half of the standard VCA-IgA kit, which contains native capsid protein and they are much easier to be got in township hospitals and institutions in China.

Reviewer 5:

1. The authors have used AUC to compare between the 7 tests with the gold standard test EBNA1-IgA. The non-inferiority margin used was 5%.They have used chi-square to test whether the

difference is within 5%. However they have not provided. This will provide us the information on what would be the maximum difference that would be obtained in the long run. May be that they should provide this using the boot-strap method.

Re: Thanks for your suggestion. In this study we use $\Delta=0.05$ as the cut point for Non-Inferiority Test for Paired ROC Curves. Here are the expiations (reference1-2 below),

Let $\theta1$ and $\theta0$ be the paired ROC curve areas for the new and the standard diagnostic kits and the Δ ($\theta0-\theta1$) be the pre-determined clinically meaningful equivalence limit. In this study, we let $\Delta=0.05$(cut point), because research always use $\Delta=0.05$ when Δ was hard to know. Then, calculated the standardized differences $\lambda1$ and $\lambda0$ for the new and the standard diagnostic kits and let $\varepsilon= \Phi-1(\theta0)-\Phi-1(\theta0+\Delta)$.The hypothesis for non-inferior test is,

H0: $\lambda1-\lambda0<\varepsilon$ versus H1: $\lambda1-\lambda0\geq\varepsilon$, $\alpha = 0.05$

If $p>0.05$, we should accept H0, which means the new kit was inferior to the standard diagnostic kit; if $p<0.05$, we should accept H1, which means the new kit was non-inferior to the standard one. We used the codes as Chen (reference3 below) and the bootstrap confidence interval (εL, εU) for the standardized difference of each new kit were:

Kits	ε	εL	εU
BB	-0.38	-0.21	0.23
BNV	-0.38	-0.24	0.16
GBI	-0.38	-0.52	-0.14
HA	-0.38	-0.31	0.04
HK	-0.38	-0.40	-0.07
KSB	-0.38	-0.13	0.27
ZS	-0.38	-0.7	-0.32

1. Obuchowski N. Testing for equivalence of diagnostic tests. Am JRadiol 1997;168:13–7.
2. Liu JP, Ma MCh, Wu ChY, et al. Tests of equivalence and noninferiority for diagnostic accuracy based on the paired areas under ROC curve. Statist Med 2006;25:1219–38.
3. Chen WZ, Zhang JY. Application of non-inferiority test for diagnostic accuracy under the areas of ROC based on bootstrap approach and its macro programming development. Modern Prev Med (Chinese) 2010;37:3009–10.

2. The table 4 analyses, it is necessary, though it comes under the broader heading of methods to study the reliability. May be this information can be made as statements with the range of values.

Re: Thanks for your suggestion. We think you mean the explanations for the range of ICC. In this study, we used guidelines for interpretation for ICC acceding to Fleiss JL (Reliability of measurement-The design and analysis of clinical experiments). The details are below:

Less than 0.40- poor;
Between 0.40 and 0.59- Fair;
Between 0.60 and 0.74- good;
Between 0.75 and 1.00- Excellent.

Since all the ICCs were larger than 0.75, we didn't write all details. We had put reference in the result part in line 39th, page 8.

3. The sub groups analyses of stage of cancer, especially Early stage (n=33) is a concern. The 95% CI for accuracy statistics are lower as compared to overall and advanced cancer. Though it is good that the authors have done this subgroup analyses, this needs to be discussed as a limitation of the tests as problem of numbers. Please provide the 95% CIs based on Bootstrap methods, for the early stage accuracies.

Re: Thank you for this good question. Due to the low percentage (less than 20%) of early stage in clinic, recruitment of early stage NPC cases is more difficult and we can only collect 33 early stage

NPC participants in our study period in Sun Yat-sen University Cancer Center (SYSUCC). We have added the comments in the Discussion, Page 12, line 14th.

The main objective of this study was to evaluate the diagnostic effects of these recombinant VCA-IgA kits and hope to provide more choices for NPC screening in the future. We found that, the EBV-related antibodies including VCA-IgA have equal diagnostic ability for early and advanced stage NPCs (reference1-2 below), and the elevated EBV antibodies even appeared several years of NPC occurrence (reference3 below). Furthermore, the data in our study also verified this point. There was no significant difference of EBV antibody level between early and late NPC patients in this study. The difference of rRod of VCA-IgA and EBNA1 for early and advanced stage NPCs were compared by Mann-Whitney U Test. The P values were 0.255 and 0.101, respectively.

The purpose of grouping was to make our study stricter but it was not the focus of our research. We used sub groups' analyses to show the sensitivities with new cutoffs for different stages and suggest that maybe there were differences between two stages of these new kits. Accurately estimation for the sensitivity of the early cases was not the core issue of this study either. So, we didn't use Bootstrap method for 95%CIs of sensitivities.

1. Stolzenberg MC, et al. Purified recombinant EBV desoxyribonuclease in serological diagnosis of nasopharyngeal carcinoma. Int J Cancer. 1996 May 3;66(3):337-41
2. Zeng Y, Zhong JM, Li LY, et al. Follow-up studies on Epstein-Barr virus IgA/VCA antibody-positive persons in Zangwu County, China. Intervirology 1983;20:190–4.
3. Cao SM, Liu Z, Jia WH, et al. Fluctuations of Epstein-Barr virus serological antibodies and risk for nasopharyngeal carcinoma: a prospective screening study with a 20-year follow-up. PLoS One 2011;6:e19100

4. The logistic regression analyses have shown a significant improvement in the AUC (that is 2 tests vs a single test). For example the AUC of KSB test increased from 0.945 (0.925, - 0.966) to 0.964 (0.947 – 0.981). The authors claim that the difference of about 1 or 2% is important? I would argue that the single test is as good as two test combined together. We need to think about the cost issues and with the trade of in the gain in accuracy. This needs to be discussed in detail. I would not go by simple significance.
Re: Thank you for this good suggestion. Generally, increasing test can improve the diagnostic efficiency but "ideal" combinations were different in different conditions. Researchers choose combinations (signal or multiple) not only by the diagnostic efficiency, but also according to financial support from community or local government, acquiring way for different kits and so on. It was really hard to analyze. But we think our research can provide more choices for NPC screening to different researchers

.

Actually, the combination of VCA-IgA and EBNA1-IgA with the "standard kits" has already been used in the NPC mass screening in part of south China and got good feedback. This combination could limit the false positives rate and the false negative rate to reasonable ranges and it could also reduce cost by reducing the number of the "high-risk" people for next examinations (nasopharyngeal fiberscope, CT) (reference below). Since the cost of recombinant VCA-IgA kits is only half of the standard one. We want to try them in the future
Liu ZW, Ji MF, Huang QH, et al. Two Epstein-Barr virus–related serologic antibody tests in nasopharyngeal carcinoma screening: results from the initial phase of a cluster randomized controlled trial in Southern China. Am J Epidemiol 2013;177:242–50.

5. The authors have used to evaluate/compare the tests with the Gold standard (GS) test. Bland and

Altman have provided warnings in using ICC in the reliability and validity studies. Please see the following REF.
A note on the use of the intraclass correlation coefficient in the evaluation of agreement between two methods of measurement. J. M. Bland and D. G. Altman.Comput. Biol. Med. Vol. 20, pp. 337-340. 1990
Re: Thanks for the suggestion. In our research, we randomly chose 10% serum samples to investigate the test-retest reliability of each kit (same method). We found the distributions of antibodies didn't conform to the normal distributions; they didn't belong to the qualitative variables either. We were afraid of that the Pearson Correlation coefficient and the Kappa analysis may not be suitable. So we decided to use the ICC at last.

We tried hard to understand the reference. It seemed that the writer thought ICC could be used as "an index of correlation between repeated measures by the same method"? But since any index has its limitations, we calculated spearman correlation coefficients indices to evaluate the test–retest reliability of these kits as follow,
The spearman Correlation Coefficients of test–retest reliabilities of
eight brands of VCA-IgA kits and the EBNA1-IgA kit

Kits	SCC	p*
VCA-IgA		
BB	0.92	<0.001
BNV	0.87	<0.001
GBI	0.85	<0.001
HA	0.87	<0.001
HK	0.86	<0.001
KSB	0.71	<0.001
ZS	0.66	<0.001
EUROIMMUN	0.82	<0.001
EBNA1-IgA	0.93	<0.001

* p<0.05 was considered as statistically significant correlation.

6. Therefore, I would ask them to provide mean and difference plots for the combination of tests compared with the GS test. They need to provide bias, 95% CI and percentage error and other related statistics meant for reliability studies. Please refer the following paper:
Bench-to-bedside review: The importance of the precision of the reference technique in method comparison studies – with specific reference to the measurement of cardiac output. Maurizio Cecconi12, Andrew Rhodes2, Jan Poloniecki3, Giorgio Della Rocca1 and R Michael Grounds2 Critical Care 2009, 13:201 (doi:10.1186/cc7129)
Re: Thank you for your suggestion. We had read the reference about Bland-Altman analysis but we were afraid that this method might not suit our objects. Because we translated the raw data (rOD, LogitP, P) into AUCs, sensitivities and specificities instead of using them directly. The raw data were only used for test–retest reliability. We calculated the spearman correlation coefficients and ICC to evaluate the test–retest reliability of these kits. But it seems the "reliability" here was not the "reliability studies" in the suggestion and reference. In our study, the "reliability" which we wanted to evaluate stood for stability of each recombinant kit. However, we showed the Spearman Correlation Coefficients between three new combinations and the standard combination and the Bland-Altman analysis between BB, HA, KSB and the standard kit below.

The Spearman Correlation Coefficients between three new combinations
and the standard combination

Combinations	SCC	p*
BB	0.92	<0.001

HA 0.94 <0.001
KSB 0.95 <0.001
* p<0.05 was considered as statistically significant correlation.

Combinations Bias with GS 95%CI Percentage error (%)
BB 0.18 -5.50-5.88 8.25
HA 0.27 -4.97-5.56 8.25
KSB 0.00 -6.35-7.51 8.00

VERSION 2 – REVIEW

REVIEWER	Minzhong Tang Wuzhou Red Cross Hospital, Wuzhou, Guangxi, P.R.China
REVIEW RETURNED	03-Feb-2017

GENERAL COMMENTS	The revised version clearly defined the research question with well written.

REVIEWER	Allan Hildesheim NCI, USA
REVIEW RETURNED	18-Feb-2017

GENERAL COMMENTS	I have reviewed responses by the authors on this manuscript. The authors have attempted to respond to the comments, but important limitations remain, including the following. 1. Limited ability to evaluate assay performance for the detection of early stage disease (N=33) 2. Inadequacy of estimates that pool across disease stage. I believe that pooling is not appropriate because the sensitivity estimates vary considerably for early and late stage disease for some of the assays (e.g., 76% vs 89% for the BB assay). The authors justify pooling based on a lack of statistically significant differences in sensitivity by stage, but this is likely driven by the small sample size (i.e., low statistical power) rather than a true lack of differences. 3. The report currently reports estimates of sensitivity and specificity based on internally optimized cutoffs only. While it is of interest to present such estimates, they should be viewed as exploratory and assay performance characteristics using a-priori manufacturer suggested cutoffs should also be presented in the manuscript. 4. Assay %CVs are not presented. The justification given is the lack of normal distribution of results. I am not sure that lack of normality precludes evaluation of %CV for assays. A statistician might be able to better comment on this point. 5. Based on clarification provided by the authors, it appears that all NPC cases were recruited from an urban area (Guangzhou) while all controls were recruited from a more rural area (Sihui) of the same province. This could have introduced biases. 6. In response to the request that information be added to Table 1 regarding the specific VCA antigen targeted by each of the assays used, the authors state that this information is proprietary and not

	available from some of the manufacturers. I understand this constraint but think that information should be provided in Table 1 when available, and that “not available from manufacturer” should be entered in the table for those assays where manufacturers refuse to provide such information. This is important to do since evaluation of results and comparison across assays is incomplete without knowing what antigen each assay is targeted. Explicitly stating in scientific presentations when manufacturers refuse to provide such information might encourage manufacturers to release such information in the future, when scientifically relevant.

REVIEWER	Zilu Zhang Harvard Medical School/Harvard Pilgrim Healthcare Institute
REVIEW RETURNED	20-Feb-2017

GENERAL COMMENTS	To whom it may concern, The authors have well addressed the problems/questions which were brought up from the first round review. The abstract is comprehensive by itself, and the article is logically consistent. The statistical figures and tables are essential and clearly presented now. The English used in te article is readable to convey the research proposal. I suggest the article may be accepted for publication without English correction.

REVIEWER	Zhigang Haung Department of Epidemiology and Health statistics,School of Public Health, Guangdong Medical University, Dongguan, P.R.China
REVIEW RETURNED	02-Mar-2017

GENERAL COMMENTS	The paper evaluated the diagnostic effects of seven Chinese recombinant VCA-IgA kits by conducting a diagnostic case-control trial with 200 cases of NPC and 200 controls from NPC-endemic areas in southern China. The results showed that three testing kits had good diagnostic accuracies and their combinations could be used in the early detection and screening for NPC, but new cutoffs need be verified in the future. The writing should be clearer and more concise but can be acceptable. Overall, this study has some practical significance and has a value for publication.

REVIEWER	Dr.L.Jeyaseelan Christian Medical College, India
REVIEW RETURNED	17-Feb-2017

GENERAL COMMENTS	The reviewer completed the checklist but made no further comments.

VERSION 2 – AUTHOR RESPONSE

Reviewer 1, 3-5: No more questions.
Reviewer 2: six questions

Reviewer 2:
1. Limited ability to evaluate assay performance for the detection of early stage disease (N=33).
Re: We did subgroup analysis and found the sensitivities were different in two stage groups and some of the differences were statistically significant. We had seen the limitation but due to the low percentage (less than 20%) of early stage in clinic, recruitment of early stage NPC cases was difficult. The phenomenon also indicated that, most patients are typically not detected until NPC was in an advanced stage. Finding out such people was also very meaningful in real life. We had added some comments in the sixth paragraph of Discussion part.

2. Inadequacy of estimates that pool across disease stage. I believe that pooling is not appropriate because the sensitivity estimates vary considerably for early and late stage disease for some of the assays (e.g., 76% vs 89% for the BB assay). The authors justify pooling based on a lack of statistically significant differences in sensitivity by stage, but this is likely driven by the small sample size (i.e., low statistical power) rather than a true lack of differences.
RE: This good question inspired us. The evaluation of sensitivity should consider the difference between two NPC stage groups. So we not only pool them together but also did subgroup analysis for each assay (table 3) and combination (table 5). As the reviewer suggested, we added some comments in the fourth and fifth paragraphs of Discussion part to emphasize the necessity of subgroup analysis. Furthermore, no differences were found between the early stage sensitivities of these three kits and that of the standard kit (0.202 for BB. 0672 for HA, 0.112 for KSB).

3. The report currently reports estimates of sensitivity and specificity based on internally optimized cutoffs only. While it is of interest to present such estimates, they should be viewed as exploratory and assay performance characteristics using a-priori manufacturer suggested cutoffs should also be presented in the manuscript.
RE: To save some space and focus on the AUC, we added the table on supplementary table 1.

Supplementary Table 1. Sensitivities and Specificities based on manufacturers' cutoffs of VCA-IgA kits and the EBNA1-IgA kit

Kits	Sensitivity (%)(95%CI)			Specificity (%) (95% CI)
	Early stage (95%CI)	Advanced stage (95%CI)	Total (95%CI)	
VCA-IgA				
BB	69.7(51.0-84.0)	85.6(80.8-90.5)	83.0 (77.8-88.2)	94.5 (91.3-97.7)
BNV	72.7(54.0-87.0)	86.8(82.1-91.5)	84.5 (79.5-89.5)	94.0 (90.7-97.3)
GBI	69.7(51.0-84.0)	71.9(65.6-78.1)	71.5 (65.2-77.8)	93.0 (89.5-96.5)
HA	90.9(76.0-98.0)	85.6(80.8-90.5)	86.5 (81.8-91.2)	87.0 (82.3-91.7)
HK	84.8(68.0-95.0)	85.0(80.1-90.0)	85.0 (80.1-89.9)	85.5 (80.6-90.4)
KSB	81.8(64.0-93.0)	77.2(71.4-83.1)	78.0 (72.3-83.7)	92.5 (88.8-96.2)
ZS	57.6(39.0-74.0)	59.3(52.5-66.1)	59.0 (52.2-65.8)	95.5 (92.6-98.4)
Euroimmun	97.0(85.0-100.0)	90.4(86.3-94.5)	91.5 (87.6-95.4)	79.5 (73.9-85.1)
EBNA1-IgA	93.9(80.0-99.0)	88.6(84.2-93.0)	89.5 (84.7-93.3)	90.0 (85.8-94.2)

4. Assay %CVs are not presented. The justification given is the lack of normal distribution of results. I am not sure that lack of normality precludes evaluation of %CV for assays. A statistician might be

able to better comment on this point.
RE: We discussed with three statisticians and thought the %CV in your suggestion meant CV of difference value (test and retest result of each assay)×100%. Results were shown in the table below. We added the table on supplementary table 2.

Supplementary Table 2. CVs of difference values of test and retest result of VCA-IgA kits and the EBNA1-IgA kit

Kits	CV
VCA-IgA	
BB	0.69
BNV	0.65
GBI	1.00
HA	0.34
HK	1.55
KSB	0.86
ZS	1.24
Euroimmun	0.63
EBNA1-IgA	0.55

5. Based on clarification provided by the authors, it appears that all NPC cases were recruited from an urban area (Guangzhou) while all controls were recruited from a more rural area (Sihui) of the same province. This could have introduced biases.
RE: We analyzed the constituent of NPC cases and found that nearly half of the cases were from rural areas (rural:urban=95:105). Though no evidence showed that there were different infection rates between rural and urban people, it might cause some other unknown bias. We added comments in Limitation part and the sixth paragraph of Discussion part.

6. In response to the request that information be added to Table 1 regarding the specific VCA antigen targeted by each of the assays used, the authors state that this information is proprietary and not available from some of the manufacturers. I understand this constraint but think that information should be provided in Table 1 when available, and that “not available from manufacturer” should be entered in the table for those assays where manufacturers refuse to provide such information. This is important to do since evaluation of results and comparison across assays is incomplete without knowing what antigen each assay is targeted. Explicitly stating in scientific presentations when manufacturers refuse to provide such information might encourage manufacturers to release such information in the future, when scientifically relevant.
RE: We contacted these companies again and added some information in the table below. Unfortunately, we still couldn’t get ingredient lists of these kits and the staffs of these companies only told us limited information.

Production information of eight VCA-IgA ELISA kits and one EBNA1-IgA kit

Abbreviation of kits	Registration certificate	viral capsid antigens type
BB	20113400638	Recombinant, p18& unknown
BNV	20113400252	Recombinant, p18
GBI	20123400802	Recombinant, p18 & p23
HA	20113400814	Recombinant, p18&unknown
HK	20123400446	Recombinant, p18&unknown
KSB	20093400720	Recombinant, p18&unknown
ZS	20123400205	Recombinant, p18 & p23
Euroimmun	20133402251	Native, mixture
EBNA1	200734011080	Recombinant, BKRF1

VERSION 2 – REVIEW

REVIEWER	Minzhong Tang Cancer Center of Wuzhou Red Cross Hospital
REVIEW RETURNED	03-Apr-2017

GENERAL COMMENTS	The authors present seven recombinant VCA-IgA kits evaluation data from high-endemic region of Southern China. The manuscript is well written, results have been clearly defined, and author address the reviewers' questions well. Overall, it is an important study, and should be considered for publication.

REVIEWER	Allan Hildesheim NCI, USA
REVIEW RETURNED	14-Apr-2017

GENERAL COMMENTS	No additional comments.

REVIEWER	Zilu Zhang Havard Medical School/Harvard Pilgrim Healthcare Institute
REVIEW RETURNED	03-Apr-2017

GENERAL COMMENTS	No more questions.

REVIEWER	Zhigang Haung Department of Epidemiology and Health Statistics, School of Public Health, Guangdong Medical University, Dongguan, P.R. China.
REVIEW RETURNED	03-Apr-2017

GENERAL COMMENTS	The authors matured the article by constant revision and the answers were persuasive. I suggest the article should be accepted for publication.