PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Tryptophan and phenylalanine are associated with arsenic–induced skin lesions in a Chinese population chronically exposed to arsenic via drinking water: a case–control study
AUTHORS	wei, yaping; Jia, Chaonan; Lan, Yuan; Hou, Xiangqing; Zuo, Jingjing; Li, Jushuang; Wang, Tao; Mao, Guangyun

VERSION 1 – REVIEW

REVIEWER	Megan Niedzwiecki Department of Environmental Medicine and Public Health, Icahn School of Mount Sinai, New York, NY, USA
REVIEW RETURNED	07-Aug-2018

GENERAL COMMENTS	In this manuscript, Wei et al. performed a metabolomics study to identify metabolites that distinguish individuals with and without arsenic-induced skin lesions in three villages in Inner Mongolia, China. They found two amino acids (tryptophan and phenylalanine) whose levels were negatively associated with skin lesion risk. This is an interesting study and would be a valuable addition to the growing literature on metabolomics and environmental exposures, but prior to possible publication, I have questions/concerns in regards to the criteria listed on the Review Checklist. Q: Is the study design appropriate to answer the research question? R: This is a relatively small study (56 matched case-control pairs). Given the high probability of chance findings in metabolomics studies, it is important to include a validation sample/cohort when identifying biomarkers for case-control discrimination, which was not present in the design of the current study. Q: Are the methods described sufficiently to allow the study to be repeated?

	R: Even though the authors cited papers throughout the methods section that describe their methods in detail, it would still be useful for the reader if the authors added considerably more detail to this section. In the Metabolomics section (Lines 61-72), it is important to have more detail on the lab methods since different protocols dramatically affect the metabolites that are detected in a particular metabolomics assay, which affects how the results should be interpreted. This is especially important because the paper that is cited for the methods (citation #24) is not yet published, so I was unable to get a full idea of what was done (e.g., preprocessing steps, analytic platform, chromatography type, solvents, etc.). In particular, it would be very useful to know how the metabolite identifications of the amino acids was achieved (MS/MS, reference standards, etc.). Further, it is unclear why the four amino acid metabolites in this current study were selected for additional analysis from the 114 metabolites in the untargeted analysis that differed between cases and controls. The aim listed at the end of the Intro states that "Based on our previous non-targeted metabolomics data, the present study aims to quantitatively examine the association of several specific AA with the risk of AISL and their ability to predict AISL"-- first, it is unclear to which previous non-targeted metabolomics data/study the authors are referring to because there is no citation (is this also citation #24?). Second, the analysis in the manuscript is seemingly presented as discovery-based using untargeted metabolomics, yet it is unclear whether or not the amino acids were the top hits from the untargeted analysis or whether the amino acids were cherry picked from the 114 discriminatory metabolites based on biological hypotheses about the roles of amino acids in skin lesion development - please explain in more detail how/why amino acids were specifically chosen from the untargeted assay for further analysis. If the current paper relies on findings from a previous study (citation #24?), please discuss this study in more detail. It does seem like this analysis from this paper (amino acids from the untargeted screen) could have been included in the previous paper rather than separated into its own paper. Q: If statistics are used are they appropriate and described fully? R: In the Statistical Analysis section (Lines 73-93), how were the confounding variables selected for inclusion in the GLMs? Also, given the case-control matching scheme, why were conditional logistic regression models not used? The results state that there was not a significant interaction between tryptophan and phenylalanine on skin lesions (p=0.46, Lines 155-156). Given this, can the authors provide justification for examining the joint effect of these two variables? Also, given the small numbers of participants and cases within each stratum, the sample size seems to be underpowered for such an analysis. Q: Are the discussion and conclusions justified by the results?

	R: Given the small sample size and lack of a validation cohort, the conclusions should be tempered, especially in Lines 247-250. It is surprising that the skin lesion cases did not have higher arsenic exposures than the controls and did not markedly differ on other demographic characteristics - it may be useful to include a brief discussion of this point. Would these similarities be expected, or was the study underpowered to detect differences? Q: Are the study limitations discussed adequately? R: The paper needs to more strongly emphasize the limitations of its small sample size and not having a validation cohort. There will also likely be limitations related to the metabolomics methods that also need to be discussed (as there are always limitations in any untargeted metabolomics study), but given the lack of information on these methods provided in the paper, it is impossible to know what needs to be discussed. Q: Is the standard of written English acceptable for publication? R: The paper would benefit from copy editing to improve grammar and flow.

REVIEWER	Margaret Karagas Geisel School of Medicine at Dartmouth, USA
REVIEW RETURNED	18-Aug-2018

GENERAL COMMENTS	This manuscript reports on a serum metabolomic analysis of amino acids from a nested case-control study of arsenic-induced skin lesions (AISL) from a cohort chronically exposed to arsenic from Inner Mongolia. Identifying metabolomic differences associated with environmentally induced diseases using emerging high through put methods has great promise. Application of this approach to understand arsenic-induced skin lesions has global significance, and the findings are potentially interesting. Specifically, the authors suggest that the biomarkers found to be inversely related to AISL could serve as biomarkers for early detection of those who might go on to more serious health sequelae. 1. An advantage of the study is that it is nested within a cohort study. This, in theory, would provide the authors the opportunity to examine arsenic exposure and serum metabolites in relation to subsequent occurrence of AISL. However, a description of the design is necessary to fully understand whether the underlying purpose of the study is simply to identify markers that distinguish the two groups, versus elucidating etiologic mechanisms and potential preventive strategies. This includes the eligibility criteria used to enlist the cohort originally (e.g., where were they recruited from e.g., clinics, village rosters etc., what were the levels of arsenic in the drinking water, were there any exclusion based on disease status (i.e., did they include prevalent AISL), what type of baseline screening was performed etc., what were the response rates, how long and how they were followed for AISL,

	and importantly at what point urine and blood were collected, including whether samples were collected pre- or post-diagnosis of AISL. Given the relatively small size of the study, these features are needed to ensure the control are representative of the population in which the cases arose (i.e., their comparability). Fundamental aspects of the study design are likewise important to interpret the study results. 2. Further, through the manuscript the authors need to make a clearer distinction between observed differences that may reflect: (1) dietary intake or exogenous exposures other than arsenic, (2) the impact of arsenic exposure on metabolic processes, and (3) metabolites related to their outcome, skin lesions. Did the authors have information about the diets of their participants? 3. The authors make statements that are not so strongly substantiated and could be misinterpreted. One example appears on lines 16-18: “Studies have shown that arsenic methylation in vivo is tightly associated with metabolic syndrome, which is believed to be related to many metabolites.” While there is evidence of a relationship between arsenic methylation capacity and metabolic syndrome, the association is not conclusive. The authors analyzed urinary arsenic in the participants. It is noteworthy that there were no differences in the AISL and control groups with respect to arsenic exposure. Can the authors explain why? Did they examine arsenic methylation capacity (i.e., ratios of the metabolites) in relation to AISL or serum AAs?

REVIEWER	Dr Francesca Chappell University of Edinburgh, United Kingdom
REVIEW RETURNED	01-Nov-2018

GENERAL COMMENTS	Please note that my review is restricted to the statistical aspects of the paper as I am not a clinician. The authors have written a paper investigating the association between metabolites and the development of arsenic-induced skin lesions. They have obviously taken a lot of care and I hope they feel that my comments will improve the paper. In no particular order: 1. The study design is case-control with 56 cases and 56 controls. The metabolites were selected via a data-driven method (page 7 lines 79 to 81, "a multiple stepwise regression analysis was applied to screen relevant AA independently associated with AISL"). Such data driven methods are notorious for selecting predictors that do not perform well in other datasets, particularly when selected from small datasets. My concern is that the metabolites selected for the analysis will not be predictive in other patients. Please read: https://www.ncbi.nlm.nih.gov/pubmed/15184705. This is a major limitation of the study.

	2. One of the aims of the data driven methods was to avoid collinearity between metabolites. I was glad to see consideration of collinearity, but data driven methods are not the answer. The authors could use the vif function available in the R package car to explore collinearity instead. Also, it seems to me that in their analyses, they used only one metabolite at a time, so collinearity between metabolites could not occur. 3. Why split the continuous data into quartiles? Where there nonlinear effects? Please justify this choice. Generally, splitting continuous data into groups results in a loss of statistical power and should be avoided. 4. Page 10, Table 3. The adjusted analyses included metabolite, age, gender, smoking, alcohol, fasting plasma glucose, triglycerides, low-density lipoprotein and % inorganic arsenic - nine predictors. Please read https://www.ncbi.nlm.nih.gov/pubmed/8970487. The rule of thumb is to have no more than 10 events per predictor, so here the maximum would be 5 as there are only 56 cases. Some statisticians regard the 10 events per variable rule of thumb to be over optimistic. Using too many predictors will again result in a statistical model that does not perform well in other datasets. 5. Do the authors have the data to compare the 450 - (2x56) = 338 people excluded from the study with the 2x56 = 112 people included in the study? I would like to see that table with age, gender, smoking status, etc. 6. Thinking about this study as a diagnostic study - the authors mention sensitivity and specificity - it would be a Phase 1 diagnostic study due to the case-control design. Phase 1 diagnostic studies are similar to feasibility studies, they tend to overestimate sensitivity and specificity and hence the area under the ROC curve. This point needs to be included as a limitation. Please read: https://www.ncbi.nlm.nih.gov/pubmed/15961549 and include this as a limitation. 7. How did they check the fit of the logistic regression models? Discrimination and calibration statistics and plots please. 8. Page 6 line 75. The authors used the Shapiro-Wilk test to look at Normality of predictors. Unfortunately, many tests for Normality are affected by sample size rather than Normality. My preferred way to check for Normality is to look at histograms (hist in R) and QQ plots (qqnorm and qqline R functions). 9. Where are the ROC curves? 10. • P.8 Table 1. Some of the p values do not seem to match the data, e.g. DMA% median and IQR is 0.62 (0.57 to 0.71) for AISL and 0.62 (0.48 to 0.65) for non-AISL, and yet the p value is quite small = 0.096. Could the authors please check this table? 11. In the abstract and other places, the authors use the words "probability" and "risk" to describe results which must be odds ratios. Odds ratios are not probabilities or risks, so could the authors please amend the manuscript to reflect this.

VERSION 1 – AUTHOR RESPONSE

Reviewer: 1

1.1 Q: Is the study design appropriate to answer the research question?

R: This is a relatively small study (56 matched case-control pairs). Given the high probability of chance findings in metabolomics studies, it is important to include a validation sample/cohort when identifying biomarkers for case-control discrimination, which was not present in the design of the current study.

Response:

We agree that the sample size (56 patients matched 56 controls) of the present study is really not too large. However, as we can see from Figure 1S, a total of 56 pairs of participants can well balance the power of tests in the current study.

Nowadays, arsenic-induced skin lesion (AISL) has been widely accepted as the major health impairment of arsenicalism in a chronic arsenic exposure population and is not completely curable but preventable. It is of great importance to perform the detection, diagnosis and provide appropriate treatment as early as possible. With the recent rapid advance of high-throughput technologies, metabolomics approach has become one of the most commonly used attractive biomarker candidate discovery tools. We also completely agree with the reviewer that the validation phase is important and necessary in identifying biomarkers based on metabolomics studies. However, in the present study, the 4 specific amino acids were extracted from 28 detected metabolites, which were identified based on both the discovery and validation phases in a previous metabolomics study of our team[3]. We mainly aimed to investigate whether the odds of AISL was independently associated with these 4 specific amino acids levels. In addition, we also want to assess whether they can be used as potential biomarkers to distinguish AISL from their counterparts early and contribute to further mechanistic studies. This is why no validation design and results could be found in our manuscript.

1.2. Q: Are the methods described sufficiently to allow the study to be repeated?

R: Even though the authors cited papers throughout the methods section that describe their methods in detail, it would still be useful for the reader if the authors added considerably more detail to this section.

In the Metabolomics section (Lines 61-72), it is important to have more detail on the lab methods since different protocols dramatically affect the metabolites that are detected in a particular metabolomics assay, which affects how the results should be interpreted. This is especially important because the paper that is cited for the methods (citation #24) is not yet published, so I was unable to get a full idea of what was done (e.g., preprocessing steps, analytic platform, chromatography type, solvents, etc.). In particular, it would be very useful to know how the metabolite identifications of the amino acids was achieved (MS/MS, reference standards, etc.).
Response:
Thanks a lot for this reminder. We have revised our manuscript as suggested. Please see details in the “UPLC-MS/MS Metabonomic Profiling” part of the method section).
Serum samples（200 μL in microcentrifuge tubes) were thawed to room temperature (25°C) and 600 μL mixture (90% acetonitrile - 10%water) were added to each sample. The samples were vigorously mixed for 20 seconds and centrifuged for 5 min at 12000 rpm (20°C). The top 400 μL of each supernatant were then transferred and dried down in a vacuum concentrator centrifuge. The dried samples were re-suspended in 130 μL of water (including 15% acetonitrile), mixed vigorously for 20 seconds and repeated the centrifugation method described above. Two μL of the supernatant were collected as samples to be determined. Serum metabolic profile acquisition was performed by using ACQUITY UPLC®/Xevo® G2 QTof/MS^{E} (Waters Corp., Milford, MA, USA). Chromatographic separation was performed at 50°C using a WATERS HSS T3 column (2.1×100 mm, 1.7 μm) with a flow rate of 0.4mL/min. The mobile phase was a mixture of (A) H_2O with 0.1% formic acid and (B) methanol with 0.1% formic acid. Elution was in linear gradient with the programmed gradient at 0 min with 100% A and 0% B, 1.00min with 100%A and 0% B, 8 min with 0%A and 100% B, 13.00 min with 0% A and 100% B. The mass spectrometer was operated under both positive-ion (ESI^+) mode and negative-ion (ESI^-) mode electrospray ionization. The scan range was from 50 to 1200 m/z. Data was collected in both ESI^+ and ESI^- modes. Capillary voltage was set at 3000 V and 2500 V, respectively. The desolvation flow rate was 800 L/h at 350°C. Argon was used as a collision gas, and the collision energy was adjusted from 10 eV to 40 eV for each analysis. Quantum clustering (QC) samples were prepared by pooling aliquots of each sample and used to reflect the reliability of further metabolomics analysis. After peak deconvolution, alignment, integration and normalization, the data including retention time (RT), mass to charge ratio(m/z), and peak intensity were extracted from raw chromatograms using Progenesis QI 2.0 (Waters Corp., Milford, MA, USA). The MS/MS mode was performed to obtain metabolites levels processed with MarkerLynx Applications Manager Version 4.1 (Waters Corp., Milford, MA, USA).

Further, it is unclear why the four amino acid metabolites in this current study were selected for additional analysis from the 114 metabolites in the untargeted analysis that differed between cases and controls. The aim listed at the end of the Intro states that "Based on our previous non-targeted metabolomics data, the present study aims to quantitatively examine the association of

several specific AA with the risk of AISL and their ability to predict AISL"-- first, it is unclear to which previous non-targeted metabolomics data/study the authors are referring to because there is no citation (is this also citation #24?).

Response:

Thank you for raising this important issue. Firstly, we are so sorry to point out that a total of 70 rather than 114 metabolites were observed to be different between the health lesions induced by chronic arsenic exposure participants due to our clerical errors. Secondly, in our previous study (citation #24), which aimed to investigate whether metabolomics approach could be applied in the early identification of the health lesions induced by chronic arsenic exposure (HLICAE) in a chronic arsenic exposure population via drinking water, we detected that the combination of 28 differential metabolites including the 4 amino acids could be effective net-work biomarkers in distinguishing HLICAE from their counterparts. However, whether amino acids metabolism independently associated the odds of arsenic-induced skin lesions (AISL) remains unclear. To answer this question, we carried out this study to quantitatively examine whether the odds of AISL were significantly associated with the 4 amino acids and the potential mechanisms of AISL development linked to amino acid metabolism in a community-based arsenic chronic exposure population. We have added essential associated contents in the revised manuscript.

Second, the analysis in the manuscript is seemingly presented as discovery-based using untargeted metabolomics, yet it is unclear whether or not the amino acids were the top hits from the untargeted analysis or whether the amino acids were cherry picked from the 114 discriminatory metabolites based on biological hypotheses about the roles of amino acids in skin lesion development - please explain in more detail how/why amino acids were specifically chosen from the untargeted assay for further analysis. If the current paper relies on findings from a previous study (citation #24?), please discuss this study in more detail. It does seem like this analysis from this paper (amino acids from the untargeted screen) could have been included in the previous paper rather than separated into its own paper.

Response:

Thank you for raising this important issue. As an attractive biomarker discovery tool, metabolomics approach has been commonly used to study the changes in the biochemical composition of biofluids, including serum, in response to adverse toxic events. Many of these in vivo studies suggested that the individual changes in metabolites and patterns of them are observed to be highly associated with the toxicological changes. However, majority of these proposed novel biomarkers are lack of sufficient specificity, which may reflect non-specific effects of a number of different types of toxicity rather than a specific pathology. Based on appropriate multivariate statistical data analysis, Connor et al[4] reported that a number of metabolites associated with energy metabolism showed obvious alterations in the urinary metabolic profiles of male Wistar Han rats. Although being considered to be non-specific markers of toxicity, these altered metabolites are also identified as essential biomarkers of a specific type of toxicity in

some instances[4-6]. As one of the important metabolic pathways in organisms, the impacts due to chronic arsenic exposure on amino acids metabolism should also be extensively assessed. Both animal and epidemiological studies revealed that chronic arsenic exposure disrupted amino acid metabolism[5 6]. These findings would be helpful on improving the identification of specific biomarkers or patterns of specific toxic effects. However, few works have been applied to comprehensively examine the independent association of arsenic-induced skin lesions (AISL) relevant to amino acids mechanism. This is why we carried out this study.

In our previous untargeted metabolomics study, we mainly focused on association between the combinations of 28 detected metabolites as a net-work biomarker with the odds of HLICAE. All of the aforementioned 28 metabolites including the 4 amino acids of the present study were separately detected based on the discovery phase and evaluated in the validation phase. Though amino acid metabolism was covered but merely a minor unit of the work, the independent association of HLICAE with amino acids were not fully examined.

In the present study, we aim to investigate the independent association between the odds of AISL and the 4 amino acids, and to assess whether they could be potential biomarkers in the early detection of AISL. Meanwhile, the outcome of the present study was AISL instead of HLICAE. Our findings suggested that the alterations of tryptophan and phenylalanine were important implications of health condition of chronic arsenic exposure population and AA metabolism might independently play an important role in AISL initiation. In addition, we have revised our manuscript accordingly as suggested.

1.3. Q: If statistics are used are they appropriate and described fully?

R: In the Statistical Analysis section (Lines 73-93), how were the confounding variables selected for inclusion in the GLMs?Also, given the case-control matching scheme, why were conditional logistic regression models not used?

Response:

Thank you for this kind reminder. As we know, logistic regression model is also one of the components of generalized linear regression models (GLMs). In the present study, conditional logistic regression models were actually carried out to assess the independent associations between the odds of AISL and the 4 amino acids. Unfortunately, we had forgotten to specify this in the manuscript because of our negligence. We have revised the manuscript as suggested.

The results state that there was not a significant interaction between tryptophan and phenylalanine on skin lesions (p=0.46, Lines 155-156). Given this, can the authors provide justification for examining the joint effect of these two variables? Also, given the small numbers of participants and cases within each stratum, the sample size seems to be underpowered for such an analysis.

Response:

Thank you for raising this important issue. When investigating the causal effects of several exposures, it is common that the effect of one exposure on the outcome will be affected by the presence or absence of other exposures. When this is the case, we say that there is interaction between the exposures. Actually, interaction is a kind of action that occur as two or more exposures have effect upon one another. So, assessing interaction between exposures can help determine which subgroups would be affected most and better understand the mechanisms of a specific disease[7]. Although not exactly the same as interaction, joint impact of two or more exposures also can help discover which subgroups would influence the outcome to the greatest extent and has been commonly used in previous studies[8 9]. In the present study, we observed that both tryptophan and phenylalanine were independently associated with the odds of AISL. Although there was no significant interaction between tryptophan and phenylalanine, obvious joint impact of these two amino acids on the odds of arsenic-induced skin lesions (AISL) was observed. Our results revealed that participants with higher levels of tryptophan and phenylalanine would have the lowest odds of AISL versus their counterparts.
To the best of our knowledge, small sample size does not mean it is insufficient. As can be seen from Figure 1S of this response, the sample size of the current study is sufficient to satisfy the statistical requirements. So, we do think that no significant interaction between the two amino acids on AISL can't be entirely attributed to insufficient sample size.

1.4. Q: Are the discussion and conclusions justified by the results?
R: Given the small sample size and lack of a validation cohort, the conclusions should be tempered, especially in Lines 247-250.
Response:
Thank you for raising this important issue. We have provided detailed explanation of the "small sample size" and information on validation in the response above. We have also revised "Taken together, two AAA's (tryptophan, phenylalanine) reduction were closely linked to the higher risk of AISL. It also suggests that tryptophan and phenylalanine are useful for distinguishing AISL earlier or screening of the high-risk individuals from their counterparts in a long-term exposed to low-level arsenic population via drinking water." as "In conclusion, the reduction of both tryptophan and phenylalanine might be independently linked to AISL and amino acids metabolism may play an important role in AISL onset."

It is surprising that the skin lesion cases did not have higher arsenic exposures than the controls and did not markedly differ on other demographic characteristics - it may be useful to include a brief discussion of this point. Would these similarities be expected, or was the study underpowered to detect differences?
Response:
We appreciate this gentle reminder by the reviewer. The probability of the initiation and development of arsenic-induced skin lesions (AISL) would be affected by a large number of

factors including arsenic exposure, metabolism, age, gender, life styles and others. In the current study, we mainly focus on investigating the association between the odds of AISL and some specific amino acids, which were determined by ultra-high-performance liquid chromatography quadrupole time-of-flight mass spectrometry (UPLC-QTOF-MS). The ability of them on the early detection of AISL was also assessed with receiver operating characteristic (ROC) analysis. So, cofactors such as age, gender, arsenic exposure, life styles, etc. would be important confounding factors and largely affect our results. To adjust for the impacts due to these cofactors, we firstly selected all participants using permuted block randomization from a single rural area in which population were chronically exposed to arsenic in a same way, had similar life style and environmental factors. Secondly, the cases and controls were matched by gender and similar age (±1 year)[10]. This study design is beneficial on discovering the independent association between AISL risk and serum amino acids, and assessing the value of AISL early identification with them. All of these may be the reason why so many potential confounders including arsenic exposure do not differ significantly between the cases and control. We have revised the manuscript as suggested (highlighted with red color in discussion).

1.5 Q: Are the study limitations discussed adequately?

R: The paper needs to more strongly emphasize the limitations of its small sample size and not having a validation cohort. There will also likely be limitations related to the metabolomics methods that also need to be discussed (as there are always limitations in any untargeted metabolomics study), but given the lack of information on these methods provided in the paper, it is impossible to know what needs to be discussed.

Response:

We appreciate this comment raised by the reviewer. Please see the detailed explanation of this question in the response above (Editor Comments 3). We think that the sample size for the present study, 56 pairs would well balance the power of tests.

We also completely agree with the reviewer that the validation phase is important and necessary in identifying biomarkers based on metabolomics studies. However, in the present study, the 4 specific amino acids were extracted from 28 detected metabolites, which were identified based on both the discovery and validation phases in a previous metabolomics study[3]. Our main aim is to investigate whether the odds of arsenic-induced skin lesions (AISL) was independently associated with these 4 specific amino acids levels. In addition, we also want to assess whether they can be used as potential biomarkers to distinguish AISL from their counterparts early and contribute to further mechanistic studies. This is why no validation design and results could be found in our manuscript (Please see details Response 1.1).

1.6. Q: Is the standard of written English acceptable for publication?

R: The paper would benefit from copy editing to improve grammar and flow.

Response:

Thank you for this valuable comment. A native English speaker has reviewed the manuscript carefully to improve grammar and flow.

Reviewer: 2

2.1. An advantage of the study is that it is nested within a cohort study. This, in theory, would provide the authors the opportunity to examine arsenic exposure and serum metabolites in relation to subsequent occurrence of AISL. However, a description of the design is necessary to fully understand whether the underlying purpose of the study is simply to identify markers that distinguish the two groups, versus elucidating etiologic mechanisms and potential preventive strategies. ①、This includes the eligibility criteria used to enlist the cohort originally (e.g., where were they recruited from e.g., clinics, village rosters etc.,

Response:

We appreciate this valuable suggestion raised by the reviewer. To make the manuscript more concise, we only described the study design briefly in the manuscript and the detailed information on the study design could be found in our previous paper[10]. In this revised manuscript, we have added these essential contents as suggested.

what were the levels of arsenic in the drinking water, were there any exclusion based on disease status (i.e., did they include prevalent AISL), what type of baseline screening was performed etc., what were the response rates, how long and how they were followed for AISL, and importantly at what point urine and blood were collected, including whether samples were collected pre- or post-diagnosis of AISL.

Response:

In general, the level of arsenic exposure can be assessed by the determinations of drinking water, urinary arsenic profiles and others. The urinary arsenic profiles concentration could be considered as the internal exposure, while the arsenic in the drinking water was thought as the external exposure. It is widely accepted that the internal exposure would be more appropriate than the external exposure when assessing the effects of a specific exposure. As the internal exposure, the urinary arsenic profiles concentration may individually account for the accurate exposure level of arsenic and are believed to be much better than other external exposures. So, in the present study, the individual arsenic exposure level was mainly determined by the urinary arsenic species profile including the trivalent arsenic, pentavalent arsenic, monomethylarsonous acid (MMA), dimethylarsenate (DMA) and others. No data about the arsenic level in the drinking water were determined in the present study as they could not represent the accurate level of arsenic exposure individually.

As specified in the manuscript, this study was originally from a randomized, double-blind, and placebo controlled clinical trial (NCT02235948), in which all subjects were randomly selected using permuted block randomization from a single rural area in a population chronically exposed

to low-level arsenic via drinking water, had indistinguishable life style and influences from very much alike other environmental factors. Information on the inclusion and exclusion criteria of the participants could be found in our previous study[10]. Strictly following the criteria of arsenicosis[11], AISL was diagnosed as the presence of arsenic induced keratosis, hyperpigmentation or depigmentation at the beginning of the trial. Urine and blood samples were also collected at the time point of participants' enrollment. Our results were mainly based on the baseline data of the trial. No information on the follow-up and its associated issues were included in the manuscript. We have revised the manuscript accordingly.

Given the relatively small size of the study, these features are needed to ensure the control are representative of the population in which the cases arose (i.e., their comparability). Fundamental aspects of the study design are likewise important to interpret the study results.

Response:

Thank you for raising this important issue. We completely agree that the issues raised by the reviewer are important to interpret our findings. Please see details in the former part of this response. We have also carefully revised the manuscript accordingly.

Specific recruitment criteria are described in our previous paper[10]. The study was carried out between September 2010 and December 2011 in a population of 3 arsenic exposed villages stratified and randomly selected, based on the results of average arsenic concentration tests in the last 2 decades in Wuyuan county of Hetao Plain, Inner Mongolia, China. Of 653 total residents in the above 3 villages, 450(men 169; women 281) residents, at age of 18 to 79 years, were recruited for a randomized clinical trial. The inclusion criteria were: men or women more than 18 years of age and chronically exposed to arsenic (arsenic concentration of the drinking water >10 mg/L), those who had no folic acid supplementation in the 2 weeks before the study, women of childbearing age agreed to use a reliable contraception method during the study, and everyone volunteered to participate and signed informed consent. Patients were excluded if they were pregnant or breast-feeding women, were allergic to folic acid, had clearly defined allergic history, reported long-term use of folic acid and other B vitamins, had obvious signs, including gastritis, ulcer, etc. or laboratory abnormalities, which could affect the efficacy of folic acid or were otherwise deemed unsuitable to participate in the study based on the judgment of the investigators, did not agree to cancel the medications, which may affect serum folate concentration during the study period. In addition, subjects who were planning to become pregnant during the study or planned to move out of the area within the study period were also excluded.

As specified in the manuscript, this study was originally from a randomized, double-blind, and placebo controlled clinical trial (NCT02235948), in which all subjects were randomly selected using permuted block randomization from a single rural area in a population chronically exposed to low-level arsenic via drinking water, had similar life style and influences under similar environmental factors. Information on the inclusion and exclusion criteria of the participants could

be found in our previous study[10]. Strictly following the criteria of arsenicosis[11], AISL was diagnosed as the presence of arsenic induced keratosis, hyperpigmentation or depigmentation at the beginning of the trial. Urine and blood samples were also collected at the time point of participants' enrollment. Our results were mainly based on the baseline data of the trial. No information on the follow-up and its associated issues were included in the manuscript. We have revised the manuscript accordingly.

2.2 Further, through the manuscript the authors need to make a clearer distinction between observed differences that may reflect: (1) dietary intake or exogenous exposures other than arsenic, (2) the impact of arsenic exposure on metabolic processes, and (3) metabolites related to their outcome, skin lesions. Did the authors have information about the diets of their participants?

Response:

Thank you for raising this important issue. In the present study, both the cases and controls were selected from a same stable rural area, chronically exposed to low-level arsenic in a same way, had had similar life style and influences under similar environmental factors. In addition, participants with AISL were also matched with the controls by gender and very similar age (±1 year), which would additionally adjust for the impact due to arsenic exposure time as they were chronically exposed to arsenic via drinking water from their birth. This perhaps was the reason why we could not detect obvious difference in demographics, clinical characteristics and urinary arsenic species (Table 1). So, we do believe that the influence of exogenous exposures, impact of arsenic exposure on metabolic process and other covariates were much more comparable between the cases and controls though we have no accurate additional information on them.

2.3 The authors make statements that are not so strongly substantiated and could be misinterpreted. One example appears on lines 16-18: "Studies have shown that arsenic methylation in vivo is tightly associated with metabolic syndrome, which is believed to be related to many metabolites." While there is evidence of a relationship between arsenic methylation capacity and metabolic syndrome, the association is not conclusive.

Response:

Thank you for raising this important issue. We have revised the sentence as "Previous studies reported that arsenic methylation in vivo might be associated with metabolic syndrome".

The authors analyzed urinary arsenic in the participants. It is noteworthy that there were no differences in the AISL and control groups with respect to arsenic exposure. Can the authors explain why? Did they examine arsenic methylation capacity (i.e., ratios of the metabolites) in relation to AISL or serum AAs?

Response:

Thank you for raising this important issue. We believe that the accurate determination of the arsenic exposure level is crucial for the following investigation. In general, urinary arsenic concentration will be considered as the internal exposure and accurately represent the individual exposure level of arsenic. Internal exposure is believed to be better than other external exposures when assessing the effects of a specific exposure. This is why we used the urinary arsenic species to account for the individual arsenic exposure level.

In the present study, all participants came from a single rural area and were chronically exposed to arsenic via drinking water. Among them, 56 participants were diagnosed with arsenic-induced skin lesion (AISL) and selected as the cases. Other 56 participants were matched with the cases by gender and similar age (±1 year) and selected as the controls. Both the cases and controls had a comparable life-style and were chronically exposed to similar arsenic levels. This perhaps was the reason why few significant differences of arsenic levels were observed in this study. In addition, we examined the relationship between arsenic methylation capacity (i.e., ratios of the metabolites) and AISL as well as these specific amino acids and added the results in the supplementary materials (Table S6, Table S7).

Table S6. Correlation matrix among urinary arsenic species and the four specific amino acids$^{\xi}$.

Urinary arsenic species profile	Tryptophan	Phenylalanine	Leucine	Phenylalanyl phenylalanine
iAS%	-0.106(0.264)	0.025(0.794)	0.055(0.562)	0.022(0.815)
MMA%	0.164(0.085)	-0.002(0.982)	0.094(0.323)	-0.053(0.577)
DMA%	-0.050(0.598)	-0.043(0.653)	-0.111(0.244)	0.004(0.965)
tAs (μg/g creatinine)	0.007(0.944)	-0.130(0.173)	-0.195(0.040)	-0.051(0.595)

ξ Data were presented as the coefficient of correlation (p-value). iAS: inorganic arsenic ($iAs^{Ⅲ}$ +$iAs^{Ⅴ}$); MMA: monomethyl arsenate ($MMA^{Ⅲ}$+$MMA^{Ⅴ}$); DMA: dimethyl arsenate ($DMA^{Ⅲ}$+$DMA^{Ⅴ}$); tAs: total arsenic ($iAs^{Ⅲ}$+$iAs^{Ⅴ}$+MMA+DMA); iAs%= iAS/tAs*100%; MMA%=MMA/tAs*100% and DMA%=DMA/tAs*100%.

Table S7 The relationship between arsenic methylation capacity and ASIL.

Arsenic profiles	β	SE	OR	95% CI		p-value
iAS%	-3.073	2.472	0.046	0.000	5.883	0.214
MMA%	-2.106	1.797	0.122	0.004	4.121	0.241
DMA%	2.556	1.517	12.884	0.659	251.973	0.092
tAs (μg/g creatinine)	-0.001	0.002	0.999	0.995	1.003	0.534

iAS: inorganic arsenic ($iAs^{Ⅲ}$+$iAs^{Ⅴ}$); MMA: monomethyl arsenate ($MMA^{Ⅲ}$+$MMA^{Ⅴ}$); DMA: dimethyl arsenate ($DMA^{Ⅲ}$+$DMA^{Ⅴ}$); tAs: total arsenic ($iAs^{Ⅲ}$+$iAs^{Ⅴ}$+MMA+DMA); iAs%= iAS/tAs*100%; MMA%=MMA/tAs*100% and DMA%=DMA/tAs*100%. β: parameters; SE: standard error; OR: odds ratio; 95% CI: 95% confidence interval;

Reviewer: 3

3.1 The study design is case-control with 56 cases and 56 controls. The metabolites were selected via a data-driven method (page 7 lines 79 to 81, "a multiple stepwise regression analysis was applied to screen relevant AA independently associated with AISL"). Such data driven methods are notorious for selecting predictors that do not perform well in other datasets, particularly when selected from small datasets. My concern is that the metabolites selected for the analysis will not be predictive in other patients. Please read: https://www.ncbi.nlm.nih.gov/pubmed/15184705. This is a major limitation of the study.

Response:

Thank you so much for this valuable advice. As we know, an outcome is usually affected by a number of factors. To examine the independent impact of a specific factor on the outcome, methods such as standardization, stratified analysis, multiple regression models and others have been developed by statisticians. Nowadays, multiple regression models including GLMs has been more and more commonly used to solve this problem. So, an appropriate model is of great importance to reveal the “real” relationship between the exposure and outcome. However, overfitting occurs frequently in a multiple regression model[12]. As contains more parameters than can be justified by the data, an overfitted model will fail to represent the real relationship of exposure and outcome or replicate in future samples, thus creating considerable uncertainty about the scientific merit of the findings to some extent[13 14]. So, it is needed and important to avoid the uncertainty due to overfitting.

The potential overfitting depends on both the number of parameters and conformability of the model structure[12]. To accurately make predictions, the commonly used approaches including stepwise regression models are developed by statisticians to select proper number of parameters in a multiple regression model[12]. Peduzzi et al.[15] suggested that the ratio of approximately 10 to 15 observations per predictor in a logistic regression model will produce reasonably stable estimations. Though stepwise regression model is not perfect, it has been widely accepted as a good option and the most commonly used approach when selecting proper parameters in a multiple regression model.

In the present study, overfitting was well considered. To make up the limitations of stepwise regression models, parameters enrolled in the conditional logistic regression models were selected by means of the results of both stepwise regression model and association of arsenic-induced skin lesion (AISL) with the amino acids assessed separately. Finally, a total of 4 covariates including body mass index, serum folate concentration, triglycerides and urinary total arsenic were selected as the major confounding factors and adjusted for when assessing the independent association between these amino acids and the odds of AISL in the revised manuscript.

3.2 One of the aims of the data driven methods was to avoid collinearity between metabolites. I was glad to see consideration of collinearity, but data driven methods are not the answer. The authors could use the vif function available in the R package car to explore collinearity instead.

Also, it seems to me that in their analyses, they used only one metabolite at a time, so collinearity between metabolites could not occur.

Response:

We appreciate this valuable suggestion. We agree with the reviewer that no collinearity will occur when assessing the association between AISL and a single amino acid at a time. However, we also want to investigate the joint impacts of several specific amino acids on the odds of AISL in the present study. In the revised manuscript, the potential collinearity among the 4 specific amino acids was assessed by means of the package "VIF" from RStudio (version 1.1.456 – © 2009-2018 RStudio, Inc.) as suggested (Table S5). Those amino acids existed obvious collinearity (variance inflation factor [VIF]>1.5) were removed from the conditional logistic regression model. This is why only two specific amino acids were enrolled in the final model. We have revised the manuscript accordingly.

Table S1. Results of collinearity assessment among the 4 specific amino acids.

Amino acids	Model 1	Model 2	Model 3	Model 4
Tryptophan	1.04	1.04	1.04	1.04
Phenylalanine	4.56		1.50	1.04
Leucine	4.21	1.39		
Phenylalanyl Phenylalanine	1.49	1.38	1.48	

Model 1: Tryptophan, Phenylalanine, Leucine and Phenylalanyl Phenylalanine

Model 2: Tryptophan, Leucine and Phenylalanyl Phenylalanine

Model 3: Tryptophan, Phenylalanine and Phenylalanyl Phenylalanine

Model 4: Tryptophan and Phenylalanine

3.3 Why split the continuous data into quartiles? Where their nonlinear effects? Please justify this choice. Generally, splitting continuous data into groups results in a loss of statistical power and should be avoided. Splitting continuous data into groups results in a loss of statistical power, especially in the case of small sample size.

Response:

Thank you so much for this advice. In many studies, the relationships between the outcome and indicators are usually estimated with traditional linear regression model, which usually assumes that the outcome linearly related to those indicators. Unfortunately, it is not always the truth. Traditional linear regression model will greatly, at least partly, affect the "real" relationship between the outcome and indicator and decrease the credible of the estimation when the relationship is non-linear. To overcome this problem, statisticians suggest that the continuous data can be classified into several categories and use dummy variables to explore the nonlinear relationships between the exposures and outcome instead of using the traditional linear regression model. Among the aforementioned categorical approaches, quartiles are more commonly used than others.

In the present study, we separately estimated the “real” relationships between the odds of AISL and each specific amino acid based on locally weighted regression models (Figure 1). The “dose-response” curves clearly revealed that majority of the relationships were obviously nonlinear. So, we split the continuous data into quartiles instead of using traditional GLMs directly to improve the robustness of inferences on the association between specific amino acids and the odds of AISL. We also agreed that splitting continuous data into groups would result in a loss of statistical power to some extent, especially in a study with insufficient sample size. However, as could be seen from the Figure 1S in this response, the sample size of the present study was sufficient enough to well balance the power of tests and the smallest number of participants among the four strata after splitting the data was 40 (20 pairs) (Table 4). Actually, in a paired design study, sample size over 40 would not be considered as too small. If the given the sample size of the present study was really insufficient, the detected joint impact of the two specific amino acids on AISL would actually be more robust than that observed in a larger study. Thus, the reviewer’s concern on the small sample size would not significantly affect the reliability of our findings.

3.4. Page 10, Table 3. The adjusted analyses included metabolite, age, gender, smoking, alcohol, fasting plasma glucose, triglycerides, low-density lipoprotein and % inorganic arsenic - nine predictors. Please read https://www.ncbi.nlm.nih.gov/pubmed/8970487. The rule of thumb is to have no more than 10 events per predictor, so here the maximum would be 5 as there are only 56 cases. Some statisticians regard the 10 events per variable rule of thumb to be over optimistic. Using too many predictors will again result in a statistical model that does not perform well in other datasets.

Response:

Thank you for raising this important issue. We have repeated the data analysis and revised the manuscript as suggested. Please see details in Table 3, Table S1, Table S2, Table S3, and Table S4.

3.5. Do the authors have the data to compare the 450 - (2x56) = 338 people excluded from the study with the 2x56 = 112 people included in the study? I would like to see that table with age, gender, smoking status, etc.

Response:

Please see details in table S8.

Table S6. Comparison of the characteristics of included and excluded participants.

Variables	Excluded	Included	P
Clinical Characteristics			
Age (years)	48.88±12.85	52.46±10.17	0.425
Exposure year (years)	44.28±13.10	47.91±11.20	0.011
Body mass index (kg/m^2)	24.40(22.20,26.70)	24.00(22.10,25.90)	0.148
Fasting plasma glucose (mmol/L)	48.88±12.85	52.46±10.17	0.004
Folate (ng/mL)	4.40(3.50,5.70)	4.20(3.30,5.20)	0.144

Total homocysteine (μmol/L)	12.11(10.25,15.58)	12.48(10.54,16.16)	0.369
Blood urea nitrogen (mmol/L)	6.20(4.93,7.53)	6.52(5.34,8.62)	0.043
Total cholesterol (mmol/L)	4.51(3.85,5.34)	4.63(4.10,5.80)	0.029
Triglycerides (mmol/L)	1.48(1.06,2.18)	1.43(1.05,1.95)	0.276
High-density lipoprotein (mmol/L)	1.11(0.91,1.31)	1.18(0.99,1.37)	0.076
Low-density lipoprotein (mmol/L)	3.02(2.55,3.52)	3.09(2.67,3.68)	0.375
Women [# (%)]	219(63.29)	62(59.62)	0.497
Cigarette smoking [# (%)]	118(35.01)	38(36.54)	0.776
Alcohol consumption [# (%)]	101(30.06)	34(33.01)	0.570
Illiteracy [# (%)]	113(32.66)	33(31.73)	0.859
Urinary arsenic species[z]			
iAS%	0.12(0.12,0.20)	0.12(0.09,0.17)	0.017
MMA%	0.26(0.16,0.28)	0.25(0.19,0.30)	0.627
DMA%	0.62(0.48,0.65)	0.62(0.49,0.70)	0.192
tAs (mg/g creatinine)	145.15(87.97,251.57)	176.04(104.41,246.73)	0.303

3.6. Thinking about this study as a diagnostic study - the authors mention sensitivity and specificity - it would be a Phase 1 diagnostic study due to the case-control design. Phase 1 diagnostic studies are similar to feasibility studies, they tend to overestimate sensitivity and specificity and hence the area under the ROC curve. This point needs to be included as a limitation. Please read: https://www.ncbi.nlm.nih.gov/pubmed/15961549 and include this as a limitation.

Response:

Many, many thanks for raising this important issue. We have thoroughly read the reference suggested by the reviewer and completely agree with the review on this topic. It is believed that diagnostic accuracy is the ability of a test to distinguish patients from some non-diseased individuals, which is usually assessed with sensitivity and specificity, and crucial for a diagnostic study. The reference standard should be the best available method to establish the presence or absence of the disease of interest[16]. In general, a diagnostic study must be evaluated by an appropriate design and in another clinically relevant population. However, our main aim is to investigate whether some specific amino acids were either independently and jointly associated with the odds of arsenic-induced skin lesion (AISL) in the present study. In addition, we also explored the early identification of AISL rather than AISL diagnosis with these amino acids. It does not mean that this is a diagnostic study. We believe that the methods used in the data analysis of the current study is sufficient and appropriate to answer our hypothesis.

To tell the truth, we have planned to perform the diagnostic study as suggested by the reviewer in the near future when having finished the enrollment of clinically relevant chronic arsenic exposure population in the same rural area. All of the findings in the present study will be evaluated. We do think the paper suggested by the reviewer is beneficial to a great extent for the appropriate study design, data collection and data analysis of our new study. Thanks again.

3.7. How did they check the fit of the logistic regression models? Discrimination and calibration statistics and plots please.

Response:

Many thanks for this reminder. We have checked the fit of the logistic regression model as suggested using the Hosmer–Lemeshow test[17] and the receiver-operating characteristic analysis. Please see details in table S8 and figure 2S.

Table S8. The results of Hosmer-Lemeshow goodness-of-fit test

	Chi-square	Df	Sig.
Model 1	4.317	8	0.827
Model 2	9.164	8	0.329
Model 3	3.714	8	0.882

Model 1: outcome=AISL; exposure=Tryptophan; cofactors=body mass index, serum folate, triglyceride and urinary total arsenic.

Model 2: outcome=AISL; exposure= Phenylalanine; cofactors=body mass index, serum folate, triglyceride and urinary total arsenic.

Model 3: outcome=AISL; exposure= Tryptophan + Phenylalanine; cofactors=body mass index, serum folate, triglyceride and urinary total arsenic.

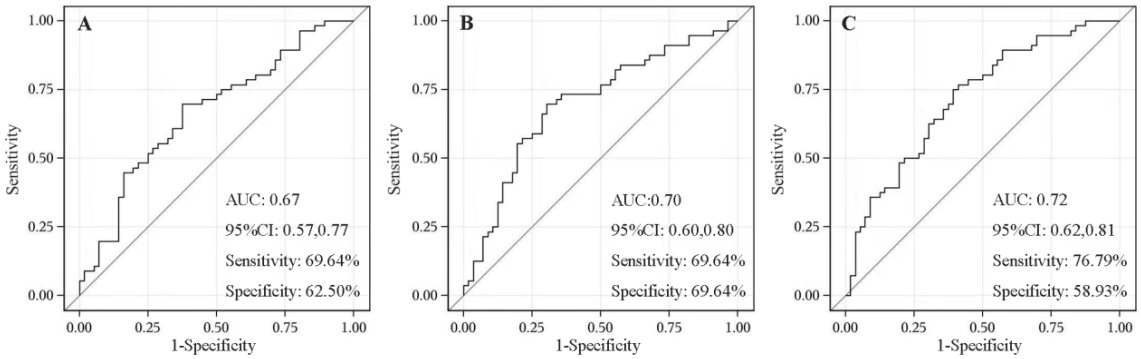

Figure 2S. The early identification value of amino acid metabolites to distinguish the participants with skin lesions induced by chronic arsenic exposure from general population via the receiver-operating characteristic (ROC) curves. A: Tryptophan; B: Phenylalanine; C: Tryptophan+ Phenylalanine

1.8 Page 6 line 75. The authors used the Shapiro-Wilk test to look at Normality of predictors. Unfortunately, many tests for Normality are affected by sample size rather than Normality. My preferred way to check for Normality is to look at histograms (hist in R) and QQ plots (qqnorm and qqline R functions).

Response:

Thank you for raising this important issue. We agree that histograms and QQ plots are commonly used approaches to observe the distribution type of the data of interest. However, they are belonging to the category of statistical description and mainly applied to represent the distribution

of a sample. As hard to avoid the subjectivity of the analyst, it is difficult to use them to determine the normality of sample instead of normality test completely[18]. Meanwhile, the Shapiro-Wilk normal test was proposed by Shapiro and Wilk in 1965 and is very commonly used nowadays, especially in study with sample size over 50[18]. Actually, we looked at normality of the 4 specific amino acids based on both QQ-plots and Shapiro-Wilk test. We have revised the manuscript accordingly.

1.9 Where are the ROC curves?
Response:
Please find the ROC curves in the above-mentioned Figure 2S.

3.10. Table 1. Some of the p values do not seem to match the data, e.g. DMA% median and IQR is 0.62 (0.57 to 0.71) for AISL and 0.62 (0.48 to 0.65) for non-AISL, and yet the p value is quite small = 0.096. Could the authors please check this table?
Response:
Thank you for this gentle reminder. As a paired design study, the comparisons of continuous variables between the cases and controls were performed with paired t-test or signed rank sum test according to their distribution types. Differences between the cases and controls among categorical variables were evaluated by McNemar-Bowker tests or Fisher's exact tests. As the distribution of DMA% was obviously skewed and the range was quite narrow (0.02-0.78 for controls vs. 0.24-0.95 for cases), it was not strange that the difference within two groups reached marginal significant level though the medians and interquartile ranges were close. We have double-checked all of the results, including table 1, of the present study as suggested.

3.11 In the abstract and other places, the authors use the words "probability" and "risk" to describe results which must be odds ratios. Odds ratios are not probabilities or risks, so could the authors please amend the manuscript to reflect this.
Response:
We appreciate this suggestion very much and have revised the associated words in the revised manuscript.

References:
1. Chen L, Cheng CY, Choi H, et al. Plasma Metabonomic Profiling of Diabetic Retinopathy. Diabetes. 2016; 65:1099-108.
2. Piszcz J, Lemancewicz D, Dudzik D, et al. Differences and similarities between LC-MS derived serum fingerprints of patients with B-cell malignancies. Electrophoresis. 2013; 34:2857-64.
3. Jia C, Wei Y, Lan Y, et al. Comprehensive Analysis of the Metabolomic Characteristics on the Health Lesions Induced by Chronic Arsenic Exposure: A Metabolomics Study. 2019.
4. Connor SC, Wu W, Sweatman BC, et al. Effects of feeding and body weight loss on the 1H-NMR-based urine metabolic profiles of male Wistar Han rats: implications for biomarker discovery. Biomarkers. 2004; 9:156-79.
5. Wang X, Mu X, Zhang J, et al. Serum metabolomics reveals that arsenic exposure disrupted lipid and amino acid metabolism in rats: a step forward in understanding chronic arsenic toxicity. Metallomics. 2015; 7:544-52.
6. Zhang J, Shen H, Xu W, et al. Urinary metabolomics revealed arsenic internal dose-related metabolic alterations: a proof-of-concept study in a Chinese male cohort. Environ Sci Technol. 2014; 48:12265-74.
7. Tyler J, Vander W, Mirjam JK. A Tutorial on Interaction. Epidemiology Methods. 2014; 3:33-72.
8. Maruyama K, Sato S, Ohira T, et al. The joint impact on being overweight of self reported behaviours of eating quickly and eating until full: cross sectional survey. BMJ. 2008; 337:a2002.
9. Phipps AI, Buchanan DD, Makar KW, et al. KRAS-mutation status in relation to colorectal cancer survival: the joint impact of correlated tumour markers. Br J Cancer. 2013; 108:1757-64.
10. Guo X, Cui H, Zhang H, et al. Protective Effect of Folic Acid on Oxidative DNA Damage: A Randomized, Double-Blind, and Placebo Controlled Clinical Trial. Medicine (Baltimore). 2015; 94:e1872.
11. Standard of diagnosis for endemic arsenism of China (WS/T211-2001),the Ministry of Health of the People's Republic of China. Beijing. 2001.
12. Harrell FEJ. Regression Modeling Strategies: Springer 2001.
13. Everitt BS, Skrondal A. The Cambridge Dictionary of Statistics: Cambrigde University Press 2010.
14. Babyak MA. What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. Psychosom Med. 2004; 66:411-21.
15. Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol. 1996; 49:1373-9.
16. Rutjes AW, Reitsma JB, Vandenbroucke JP, et al. Case-control and two-gate designs in diagnostic accuracy studies. Clin Chem. 2005; 51:1335-41.
17. Nattino G, Finazzi S, Bertolini G. A new test and graphical tool to assess the goodness of fit of logistic regression models. Stat Med. 2016; 35:709-20.
18. Liang X. Normal test. Beijing: China statistical press 1997.

VERSION 2 – REVIEW

REVIEWER	Megan Niedzwiecki Icahn School of Medicine at Mount Sinai, New York, NY, USA
REVIEW RETURNED	08-Jan-2019

GENERAL COMMENTS	The authors responded adequately to the reviewer comments. However, the paper would benefit from additional copy editing prior to publication.

REVIEWER	Francesca Chappell University of Edinburgh, UK
REVIEW RETURNED	26-Dec-2018

GENERAL COMMENTS	The authors have rewritten the manuscript and I appreciate the changes they have made. They have done a lot of work on the paper and I hope they will soon be rewarded with publication.I still have a few comments on the paper, which I hope the authors will find useful. I am unsure exactly how the authors chose a final five predictors for the model (though I am glad it is only five). If they could add a sentence or two on the criteria they used if would be helpful. Could they also please explain their decision to use quartiles for the continuous variables, Was it because of nonlinear relationships shown in Figure 1? Or was it something to do with their statement about the lack of accuracy of metabolite measurement? The authors might find this useful: https://uk.sagepub.com/sites/default/files/upm-binaries/61117_Chapter_7.pdf I couldn't find a STROBE checklist. In the discussion, a little more could be added on the limitation of a small sample and a large number of predictors.

VERSION 2 – AUTHOR RESPONSE

Reviewer: 3

I am unsure exactly how the authors chose a final five predictors for the model (though I am glad it is only five). If they could add a sentence or two on the criteria they used if would be helpful.

Response:

Thanks a lot for this reminder. To the best of our knowledge, potential confounders enrolled in a multiple regression model would be determined in the following ways: based on the comparisons of potential confounding factors between or among different groups which would be presented in the table 1 of the manuscript, screened by a stepwise regression model in which no obvious collinarity could exist and others. In the present study, we performed the predictors enrolled in our models in the two above-mentioned ways. First, we selected fasting plasma glucose (FPG), low-density lipoprotein(LDL), triglyceride(TG), urinary inorganic arsenic (iAS) and dimethyl arsenate (DMA) based on the results of the table 1 in the main text as their p-value were all less than 0.2, which was suggested as the criteria in a study with small sample size. Second, the above-mentioned 5 potential confounders was also examined using a stepwise regression model and variance inflation factor (VIF) was used to assess the potential collinarity among them. As the collinarity between %iAS and %DMA was observed and the VIF of %iAS was larger than that of %DMA, we removed %iAS from the model and selected FPG, LDL, TG and %DMA as the final potential confounders. We have added the following sentence “Variables in table 1 with p-value less than 0.2 was first selected as potential confounding factors. Potential collinarities among them were then examined in a stepwise regression model with variance inflation factor (VIF) assessment. As the collinarity between %iAS and %DMA was observed and the VIF of %iAS was larger than that of %DMA, we removed %iAS from the model and selected FPG, LDL, TG and %DMA as the final potential confounders.” Please see detail in the 3rd paragraph in the “Results” section.

1. Could they also please explain their decision to use quartiles for the continuous variables, Was it because of nonlinear relationships shown in Figure 1? Or was it something to do with their statement about the lack of accuracy of metabolite measurement? The authors might find this useful: https://uk.sagepub.com/sites/default/files/upm-binaries/61117_Chapter_7.pdf

Response:

Thank you for raising this important issue. To the best of our knowledge, there are several reasons why continuous data should be grouped[1]. Though limitations such as potential information loss and statistical power reduction especially when dichotomization is used[2,3], categories may be helpful to show a dose-response relationship, present all variables in a similar style and simplify the analysis to avoid an assumption of linearity. Investigators may choose cut-points for groupings based on commonly used values that are relevant for diagnosis or prognosis, for practicality, or on statistical grounds. They may choose equal numbers of individuals in each group using quantiles[4]. In the present study, we comprehensively examined the associations between the odds of AISL and exposures (the 4 differential AAs) in the following ways: ①using LOESS models to robustly estimate the "real" relationships between the odds of AISL and each specific amino acid based on (Figure 1); ②with the exposure as a categorical variable (quartiles) because the "dose-response" curves in figure 1 were obviously nonlinear; ③ with the exposure as a continuous variable [scaled to interquartile range (IQR)]; ④ with the exposure as a dichotomous variable which was classified by the cut-off value based on ROC analysis.

2. I couldn't find a STROBE checklist.

Response:

Thanks a lot for this reminder. Please see detail in the attached STROBE checklist.

3. In the discussion, a little more could be added on the limitation of a small sample and a large number of predictors.

Response:

Thanks a lot for this reminding. We have added "Furthermore, as it is suggested that the ratio of approximately 10 to 15 observations per predictor in a logistic regression model will produce reasonably stable estimations[5], we selected only 4 covariates in the models due to the small sample size and large number of predictors to obtain a more stable estimation." in the latter part of "Discussion" as suggested.

Reviewer: 1

The authors responded adequately to the reviewer comments. However, the paper would benefit from additional copy editing prior to publication.

Response:

Thank you for this kind reminding. We have finished the additional copy editing as suggested.

1. Barrio I, Arostegui I, Rodriguez-Alvarez MX, Quintana JM. A new approach to categorising continuous variables in prediction models: Proposal and validation. Stat Methods Med Res 2017; 26(6): 2586-602.

2. Royston P, Altman DG, Sauerbrei W. Dichotomizing continuous predictors in multiple regression: a bad idea. Stat Med 2006; 25(1): 127-41.

3. MacCallum RC, Zhang S, Preacher KJ, Rucker DD. On the practice of dichotomization of quantitative variables. Psychol Methods 2002; 7(1): 19-40.

4. Clayton D, Hills M. Models for Dose-response (Chapter 25). Statistical Models in Epidemiology; 1993.

5. Peduzzi P, Concato J, Kemper E, Holford TR, Feinstein AR. A simulation study of the number of events per variable in logistic regression analysis. J Clin Epidemiol 1996; 49(12): 1373-9.

VERSION 3 – REVIEW

REVIEWER	Francesca Chappell University of Edinburgh, United Kingdom
REVIEW RETURNED	15-Mar-2019

GENERAL COMMENTS	I'd like to thank the authors for all the extra work they have done to improve the paper. I have a couple more points to make re the selection of variables. The authors say they used stepwise selection and the criterion of p<0.2. Unfortunately such methods of variable selection can lead to spurious findings. Please read https://people.duke.edu/~mababyak/papers/babyakregression.pdf for a longer explanation of the issues. The aim of the authors was to check that the cases and controls were similar apart from AISLs, and to account for differences in the analysis - this is laudable, but I would not have used their method. I would have preferred selection of variables based on the results of previous, independent research. Could the authors please add something to the Limitations section of the paper to say that these results need to be confirmed in new studies (something which I think they would accept due to the sample size anyway), as stepwise selection methods can yield unreproducible findings. The authors have given a description of variable selection from those presented in Table 1, that is, demographic and clinical characteristics, and arsenic species. Could the authors please also explain how the four amino acid metabolites were chosen from the 70 identified?

VERSION 3 – AUTHOR RESPONSE

The authors say they used stepwise selection and the criterion of p<0.2. Unfortunately, such methods of variable selection can lead to spurious findings. Please read https://people.duke.edu/~mababyak/papers/babyakregression.pdf for a longer explanation of the issues. The aim of the authors was to check that the cases and controls were similar apart from AISLs, and to account for differences in the analysis - this is laudable, but I would not have used their method. I would have preferred selection of variables based on the results of previous, independent research. Could the authors please add something to the Limitations section of the paper to say that

these results need to be confirmed in new studies (something which I think they would accept due to the sample size anyway), as stepwise selection methods can yield unreproducible findings.

Response:
Thanks a lot for this reminder and we completely agree with the reviewer's viewpoint on the stepwise selection methods. As few similar previous, independent study could be found, the potential confounders included in the multivariable conditional logistic regression models were mainly determined when variables with p-value less than 0.2 in table 1 instead of stepwise selection. This approach had been commonly performed in many studies especially when the sample size was not too large (statistical analysis section, line 115-117). We also added a sentence "these results need to be confirmed in new studies" in the limitation section as suggested (line 281-282).
Although automated stepwise regression models have been widely utilized to screen some potential confounders, simulation studies have suggested that automated selection, unless special corrections are made, will lead to the problem of overfitting[1]. To overcome this limitation, Tibshirani and colleagues[2] developed lasso regression model to correct the traditional automated selection algorithms. This appropriate correction has been used in many researches. In the present study, glmnet package in R software, established for lasso method by statistician, was conducted firstly to screen potential confounders. We selected the years of arsenic exposure, serum folate, fasting plasma glucose (FPG), serum triglycerides (TG), low-density lipoprotein (LDL), education levels, alcohol consumption, total arsenic in urine, the percent of inorganic arsenic (%iAS), monomethyl arsenate (%MMA) and dimethyl arsenate (%DMA) based on the results of the table 1 in the main text as their p-value were all less than 0.5 and none of variables mentioned above were screened, which might be due to small effects or insufficient sample size. So, variables with p-value less than 0.2 in the comparison between two groups were selected as potential confounders as the sample size of the current study was not too large. As too many covariates would lead to overfitting to some extent[3], we finally selected 4 of them to avoid overfitting when quantifying the association between amino acids and arsenic-induced skin lesions.
Furthermore, as the distinct metabolites might be high related to each other, collinearity should also be well considered. VIF package in R software was applied to detect the potential collinearity, which is assessed by variance inflation factor (VIF), among them. When VIF is greater than 2, it was considered as collinearity existing and the associated variables were removed from the model. Finally, FPG, LDL, TG and %DMA were selected as the final potential confounders based on the results of lasso regression model combined with VIF assessment.

Could the authors please also explain how the four amino acid metabolites were chosen from the 70 identified?
Response:
Thanks a lot for this reminding. In the present study, we mainly focus on investigating the association of specific serum amino acids (AAs) with the odds of arsenic-induced skin lesions (AISL) and their ability to distinguish AISL from the counterparts. Among the 70 distinct metabolites identified in the metabolomics approach [variable importance in the project (VIP) scores >1 in partial least-squares discriminant analysis (PLS-DA) and false discovery rate (FDR) adjusted p-value < 0.05], only 4 of them are AAs (Phenylalanine, Tryptophan, Leucine, Phenylalanylphenylalanine). This is why these 4 AAs were included in the present study (please see detail in "Distinct Metabolites Identification" of the "Methods" section, line 92-103).

[1] Altman DG, Andersen PK. Bootstrap investigation of the stability of a Cox regression model. Statistics in medicine 1989;8(7):771-83.
[2] Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society 1996;58(1):267-288.
[3] Peduzzi P, Concato J, Kemper E, et al. A simulation study of the number of events per variable in logistic regression analysis. Journal of clinical epidemiology 1996;49(12):1373-9.