PEER REVIEW HISTORY

BMJ Open publishes all reviews undertaken for accepted manuscripts. Reviewers are asked to complete a checklist review form (**http://bmjopen.bmj.com/site/about/resources/checklist.pdf**) and are provided with free text boxes to elaborate on their assessment. These free text comments are reproduced below.

ARTICLE DETAILS

TITLE (PROVISIONAL)	Research protocol for an exploratory study: VOC biomarkers identification and predictive model construction for lung cancer based on exhaled breath analysis
AUTHORS	Li, Wenwen; Dai, Wei; Liu, Mingxin; Long, Yijing; Wang, Chunyan; Xie, Shaohua; Liu, Yuanling; Zhang, Yinchenxi; Shi, Qiuling; Peng, Xiaoqin; Liu, Yifeng; Li, Qiang; Duan, Yixiang

VERSION 1 – REVIEW

REVIEWER	Federica Bianchi University of Parma, Italy
REVIEW RETURNED	20-Dec-2018

GENERAL COMMENTS	In my opinion the protocol should be modified by indicating the time required between sampling and the analysis of the Tedlar bags. This is an important issue since Tedlar bags could be permeable after some time and the VOCs could be loosen. Another important issue that should be better addressed is the way of sampling: more information about the time elapsed between lunch and the collection of the breath should be given

REVIEWER	Pouline M P van Oort Academic Medical Centre Amsterdam, the Netherlands
REVIEW RETURNED	18-Jan-2019

GENERAL COMMENTS	Review Manuscript ID: bmjopen-2018-028448 entitled "Research protocol for an exploratory study: VOC biomarkers identification and diagnostic model construction for lung cancer based on exhaled breath analysis" COMMENTS TO THE AUTHOR Thank you for asking me to review this manuscript which describes a study protocol involving the investigation of the use of exhaled breath analysis for the diagnosis of lung cancer. The field of breathomics is highly interesting and might become of great importance for the diagnosis of a whole variety of clinical diseases in the future. The manuscript of the study protocol is well-written and its potential value for readers of BMJ Open is clear to me. My comments are as follows: 1. The well-written introduction section ends with the study aims. Page 5 reads: 'We aim (…) and establish a diagnostic model

for lung cancer prediction'. To me it does not become clear how this diagnostic model is going to predict lung cancer. Be careful to mix/combine diagnostic and prognostic here. Can you please explain how the VOCs that you probably have been able to identify are enabling prediction of lung cancer?

2. Following the STROBE checklist (https://www.elsevier.com/__data/promis_misc/ISSM_STROBE_Checklist.pdf) , are you able to add any prespecified hypotheses to the last paragraph of the introduction section?

3. The inclusion criteria involve "lung cancer patients and healthy subjects aged 50 to 74". Could you please clarify:

a. The lung cancer patients group: is this a consecutive sample of subjects? How are the 389 patients 'chosen'? Are they all new patients who just present themselves to your clinic?
b. The sentence can be interpreted in two ways: are healthy subjects aged 50 to 74, or are lung cancer patients and healthy subjects aged 50 to 74?

4. Page 5, lines 39-40: "Lung cancer patients (…) and should not receive any treatment". Please specify which treatment you refer to. E.g. only cancer treatment?

5. Page 5, lines 43-44: "Healthy subjects should be negative of lung cancer on chest CT". Do all 389 healthy subjects receive chest CT? Is this part of standard care? If not, do you have ethical approval for this?

6. The reference standard for the definite lung cancer diagnosis has not been mentioned. Please formulate clearly the logistics and order of taking the breath samples and performing the additional tests/imaging which are part of the standard care to obtain lung cancer diagnosis at your centre.

7. Page 6, lines 23-26: "Breath samples will be (…) until analysis".

a. How long are they going to be stored until analysis?
b. Is there evidence that the VOCs will remain detectable for that amount of storage time?

8. Could you please describe the logistics of:
a. The transferral of the breath from the Bio-VOC to the Tedlar bag.
b. How SPME preconcentrates the breath in the Tedlar bag. I've never used the SPME technique myself, so it would be helpful to get a general idea of this procedure. What happens to the air in the Tedlar bag?

9. When it comes to the statistical analysis of volatile biomarkers, principal componant analysis (PCA) followed by logistic regression analysis is often used to reduce high dimensional datasets with VOCs. Subsequently sparse partial least square discriminant analysis (sPLSDA) with leave-one-out-crossvalidation can be used to identify the most discriminatory compounds. It is recommended to calculate correct classification rates (CCRs) afterwards (Westerhuis JA, Hoefsloot HCJ, Smit S, et al.: Assessment of PLSDA cross validation. Metabolomics 2008; 4:81–

	89). I would advice the authors to investigate whether the aformentioned analyses might enhance their statistical analysis plan. 10. Could you think of any potential confounders and is there a way to correct for these? 11. The supplementary material mentions the cleaning of the Bio-VOC syringe before next usage. Could you provide information which detergent you intend to use for this and has this been tested for the abundance of VOCs? In my experience certain detergents give off a fair amount of VOCs which can contaminate your next breath sample. 12. I applaud the strategy of simultaneously measuring the VOCs in the ambient air with each batch of breath samples. Page 7 lines 38-42 mention the exlusion of VOCs that are measured in breath samples as well as in ambient room air. I agree with this, and would only like to add the advice to mention these contaminating VOCs for instance in a table when you get to publishing your study results. It might give an interesting insight of which VOCs are likely to contaminate your GC-MS results in a similar setting.

VERSION 1 – AUTHOR RESPONSE

Replies to Reviewer 1 Prof. Federica Bianchi

1. In my opinion the protocol should be modified by indicating the time required between sampling and the analysis of the Tedlar bags. This is an important issue since Tedlar bags could be permeable after some time and the VOCs could be loosen.

Response: Thanks for pointing out this matter. It is our negligence that we do not present this issue in the original manuscript. Actually, we have assessed the storage stability of VOCs in Tedlar bag. Six VOCs, which were detected in exhaled breath, including isopropanol, n-butanol, n-heptanol, n-hexanal, n-heptanal, and n-decanal (100 ppbv for each VOC) were stored in the Tedlar bags at -40 °C. Three replicates of the VOCs samples were analyzed every day until the seventh day. The results are shown in supplementary Figure 1, which indicate a good stability of the six VOCs stored in Tedlar bags at -40 °C within seven days. Therefore, the exhaled samples were considered stable within seven days in Tedlar bags at -40 °C. In this study, most of the exhaled samples were analyzed within five days after collection, and only a few samples were analyzed until the sixth or seventh day. Therefore, a statement regarding the time between sampling and the analysis has been added in the revised manuscript (Page 6, Lines 24-27).

2. Another important issue that should be better addressed is the way of sampling: more information about the time elapsed between lunch and the collection of the breath should be given.

Response: Thank you for your thoughtful suggestion. In this study, the sampling is performed in the morning from 7:00 am to 9:00 am, and the fasting time required for breath gas sampling is 8 hours. Therefore, fasting (except drinking water) from 23:00 pm the day before sampling can meet the requirement. The correction has been made in the revised manuscript (Page 6, Lines 10-11).

Replies to Reviewer 2 Prof. Pouline M P van Oort

1. The well-written introduction section ends with the study aims. Page 5 reads: 'We aim (…) and establish a diagnostic model for lung cancer prediction'. To me it does not become clear how this diagnostic model is going to predict lung cancer. Be careful to mix/combine diagnostic and prognostic here. Can you please explain how the VOCs that you probably have been able to identify are enabling prediction of lung cancer?
   Response: "A diagnostic model" here actually refers to a predictive model, which is constructed based on exhaled VOCs biomarkers. Exhaled biomarkers were selected based on breathomics data from lung cancer patients and lung cancer high-risk people (healthy controls). Therefore, the biomarkers can reflect the potential development of lung cancer, and predict lung cancer through a combined predictive model. The correction has been made in the revised manuscript (Page 5, Lines 7-8).

2. Following the STROBE checklist (https://www.elsevier.com/__data/promis_misc/ISSM_STROBE_Checklist.pdf) , are you able to add any prespecified hypotheses to the last paragraph of the introduction section?

Response: Thank you for your reminder. The prespecified hypothesis of "the predictive model will reach 80% sensitivity and 80% specificity through cross validation" has been added in the manuscript (Page 5, Lines 8-9).

3. The inclusion criteria involve "lung cancer patients and healthy subjects aged 50 to 74". Could you please clarify:
   a. The lung cancer patients group: is this a consecutive sample of subjects? How are the 389 patients 'chosen'? Are they all new patients who just present themselves to your clinic?
   b. The sentence can be interpreted in two ways: are healthy subjects aged 50 to 74, or are lung cancer patients and healthy subjects aged 50 to 74?

Response: (a) All the patients recruited in this study were newly diagnosed lung cancer patients from two in-patient departments of Sichuan Cancer Hospital. The lung cancer patients group is a consecutive sample of subjects. However, not all hospitalized lung cancer patients in the hospital participated in the study. (b) Thanks for your reminder. Lung cancer patients and healthy subjects were both aged from 50 to 74 years. The correction has been made in the revised manuscript (Page 5, Lines 22-23).

4. Page 5, lines 39-40: "Lung cancer patients (…) and should not receive any treatment". Please specify which treatment you refer to. E.g. only cancer treatment?

Response: Thanks for pointing out this issue. The correction has been made using "should not receive any cancer treatment" in the revised manuscript (Page 5, Line 26).

5. Page 5, lines 43-44: "Healthy subjects should be negative of lung cancer on chest CT". Do all 389 healthy subjects receive chest CT? Is this part of standard care? If not, do you have ethical approval for this?

Response: Thanks for pointing out this issue. Yes, all 389 healthy subjects received chest CT before enrolment, because we enrolled healthy subjects at high-risk for lung cancer at two screening centers. These subjects have undergone annual lung cancer screening with low-dose computed tomography. This screening plan is a government financial transfer payment project from 2014, and it is not a research. So, there is no ethical approval for this screening project. However, all subjects gave written informed consent before screening and we already have ethical approval from the Ethics Committee of Sichuan Cancer Hospital relating to this study.

6. The reference standard for the definite lung cancer diagnosis has not been mentioned. Please formulate clearly the logistics and order of taking the breath samples and performing the additional tests/imaging which are part of the standard care to obtain lung cancer diagnosis at your centre.

Response: Thanks for your constructive advice. The correction has been made in the revised manuscript (Page 5, Lines 24-26). The definite pathologic diagnosis of lung cancer was based on the 2015 World Health Organization Classification of lung tumors [1]. The pathologic stages were based on the eighth edition of the TNM classification for lung cancer [2]. For patients who will not receive surgery, breath samples collection will be performed after pathologic diagnosis and prior to cancer treatment. For patients undergoing surgery, breath samples will be collected the day before the surgery. If the postoperative pathology is not primary lung cancer, the patient will be excluded.

[1] Travis WD, Brambilla E, Nicholson AG, et al. The 2015 World Health Organization Classification of Lung Tumors. Impact of Genetic, Clinical and Radiologic Advances since the 2004 Classification. J Thoracic Oncol 2015;10:1243–1260.

[2] Goldstraw P, Chansky K, Crowley J, et al. The IASLC lung cancer staging project: proposals for revision of the TNM stage groupings in the forthcoming (eighth) edition of the TNM classification for lung cancer. J Thorac Oncol, 2016, 11(1): 39-51.

7. Page 6, lines 23-26: “Breath samples will be (…) until analysis”.
a. How long are they going to be stored until analysis?
b. Is there evidence that the VOCs will remain detectable for that amount of storage time?
Response: (a) Most of the breath samples will be analyzed within five days, only a few samples will be analyzed within seven days until the sixth or seventh day. (b) The storage stability of VOCs in Tedlar bags at -40 °C has been assessed and the results are shown in the supplementary. Six VOCs, including isopropanol, n-butanol, n-heptanol, n-hexanal, n-heptanal, and n-decanal (100ppbv for each VOC), all indicated good storage stability in Tedlar bags at -40 °C within seven days. Therefore, in this study, the breath samples are considered stable at -40 °C within seven days in Tedlar bags. The correction has been made in the revised manuscript (Page 6, Lines 24-27).

8. Could you please describe the logistics of:
a. The transferal of the breath from the Bio-VOC to the Tedlar bag.
b. How SPME preconcentrates the breath in the Tedlar bag. I’ve never used the SPME technique myself, so it would be helpful to get a general idea of this procedure. What happens to the air in the Tedlar bag?
Response: (a) A three-way valve connected Bio-VOC, Tedlar bag valve, and atmosphere valve. When collecting exhaled breath, the three-way valve keeps Bio-VOC and atmosphere valve connected with Tedlar bag valve closed, to exclude the dead space gas(as shown in Figure 1(a)). After a total exhalation, the three-way valve keeps Bio-VOC and Tedlar bag valve connected, with atmosphere valve closed, to transfer breath gas into Tedlar bag (as shown in Figure 1(b)). (b) SPME mainly analyzes organic compounds based on the similarity-intermiscibility theory, which extracts organic compounds from the sample matrix on the SPME fiber until partition equilibrium is achieved between the sample matrix and the fiber coating. In practical use, it is not necessary to reach a distribution equilibrium as long as a reliable and stable linear relationship between the response value and actual concentration is obtained under strict experimental conditions. In this study, the extraction time has been optimized as 30 min to obtain as much information as possible in the shortest time. Air components such as oxygen, nitrogen, etc., may be adsorbed onto the fiber during the initial stage of extraction due to their small molecular weight. As the extraction proceeds, the organic compounds will compete for adsorption onto the fiber based on the similarity-intermiscibility theory, and then replace the air components.

(a) (b)

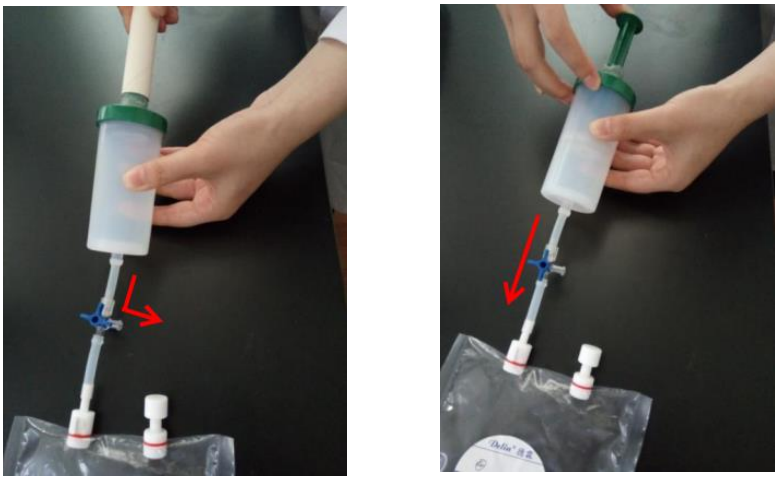

Figure 1. (a) When a subject exhales into the Bio-VOC syringe, connect the Bio-VOC valve to the atmosphere valve and close the Tedlar bag valve to remove the dead space gas. (b) When a subject stops exhaling, connect the Bio-VOC valve to the Tedlar bag valve and close the atmosphere valve to transfer the exhaled gas from Bio-VOC to the Tedlar bag.

9. When it comes to the statistical analysis of volatile biomarkers, principal component analysis (PCA) followed by logistic regression analysis is often used to reduce high dimensional datasets with VOCs. Subsequently sparse partial least square discriminant analysis (sPLSDA) with leave-one-out-cross validation can be used to identify the most discriminatory compounds. It is recommended to calculate correct classification rates (CCRs) afterwards (Westerhuis JA, Hoefsloot HCJ, Smit S, et al.: Assessment of PLSDA cross validation. Metabolomics 2008; 4:81–89). I would advise the authors to investigate whether the aformentioned analyses might enhance their statistical analysis plan.

Response: Thanks for your constructive advice. When analyzing the breathomics data, we will assess different dimensionality reduction method including PCA, linear discriminant analysis (LDA), independent component analysis (ICA), etc. Then PLSDA, sPLSDA, or OPLSDA, combined with leave-one-out cross validation (or 10-fold cross validation) will be used to identify exhaled biomarkers. The correction has been made in the revised manuscript (Page 9, Lines 9-19).

10. Could you think of any potential confounders and is there a way to correct for these?

Response: Age, sex, body mass index (BMI), smoking status (pack-years), alcohol drinking, etc. may influence the exhaled VOCs. In this study, high-risk people (healthy control) and lung cancer patients are all aged from 50 to 74 years, which reduce the impact of age. In addition, sex (male, female), BMI, smoking status (none, light, moderate, heavy, ex-smoker), and alcohol drinking history (yes, no, stop drinking) have been recorded for each subject. Univariant analysis will be used to investigate the influence of above confounders on each identified exhaled biomarker. For factors with significant impact, the correction parameter of this factor will be added to the predictive model. The correction has been made in the revised manuscript (Page 9, Lines 20-21).

11. The supplementary material mentions the cleaning of the Bio-VOC syringe before next usage. Could you provide information which detergent you intend to use for this and has this been tested for the abundance of VOCs? In my experience certain detergents give off a fair amount of VOCs which can contaminate your next breath sample.

Response: Before the next use, the Bio-VOC syringe will be cleaned with a dry dust-free cloth. There will be no detergent used. The correction has been made in the supplementary Table 1.

12. I applaud the strategy of simultaneously measuring the VOCs in the ambient air with each batch of breath samples. Page 7 lines 38-42 mention the exlusion of VOCs that are measured in breath samples as well as in ambient room air. I agree with this, and would only like to add the advice to mention these contaminating VOCs for instance in a table when you get to publishing your study results. It might give an interesting insight of which VOCs are likely to contaminate your GC-MS results in a similar setting.

Response: Thanks for your good suggestion. We will pay particular attention to the contaminating VOCs in the ambient air, and will summarize these VOCs in the experimental paper.

VERSION 2 – REVIEW

REVIEWER	Federica Bianchi University of Parma, Department of Chemistry, Life Sciences and Environmental sustainability
REVIEW RETURNED	08-Mar-2019

GENERAL COMMENTS	Nothing to declare, the manuscript has been revised according to the comments of the referees

REVIEWER	Pouline van Oort Amsterdam UMC - AMC, Amsterdam, the Netherlands
REVIEW RETURNED	28-Feb-2019

GENERAL COMMENTS	I'd like to congratulate the authors with the improvement of the paper in this revised version. It looks like a promising study and I'm looking forward to the results. Just one of my previous questions remains unanswered. Under the 'Design' header, the paper reads: “Healthy subjects should be negative of lung cancer on chest CT”. Do all 389 healthy subjects receive chest CT? Is this part of standard care? I can't imagine that you would let all healthy volunteers undergo chest CT just for the sake of this study, or do you? If so: do you have ethical approval for this? Many thanks for asking me to review this interesting paper.

VERSION 2 – AUTHOR RESPONSE

Replies to Reviewer 2 Prof. Pouline M P van Oort

1.“Healthy subjects should be negative of lung cancer on chest CT”. Do all 389 healthy subjects receive chest CT? Is this part of standard care? If not, do you have ethical approval for this?

Response: Thanks for pointing out this issue again. We are very sorry that we did not clarify this issue.

First, all healthy subjects were recruited from two lung cancer screening centres in this study. These healthy subjects were screened for lung cancer annually with low-dose CT base on an Early Diagnosis and Early Treatment of Rural Cancer Project in China from 2014. All healthy subjects have already undergone CT screening for three years at least and will undergo CT screening annually in the future. Therefore, all healthy subjects were negative of lung cancer on chest CT before enrollment. So we did not let all healthy volunteers undergo chest CT just for the sake of this study.

Second, in this study, we did not have ethical approval specifically for the CT screening again, because all the healthy subjects have already undergone CT screening annually for lung cancer which was not for the sake of this study. However, we have the ethical approval for this study from our hospital Ethics Committee (No.SCCHEC-02-2017-011).

A statement about this issue “Healthy subjects recruited from two lung cancer screening centres should be negative of lung cancer on chest CT based on a previous project “Early Diagnosis and Early Treatment of Rural Cancer” in China from 2014” has been added in the revised manuscript (page 6, lines 1-3).