PEER REVIEW HISTORY

ARTICLE DETAILS

TITLE (PROVISIONAL)	Air pollutants and development of interstitial lung disease in patients with connective tissue disease: a population-based case-control study in Taiwan
AUTHORS	Chen, Hsin-Hua; Yong, You-Ming; Lin, Ching-Heng; Chen, Yi-Hsing; Chen, Der-Yuan; Ying, Jia-Ching; Chao, Wen-Cheng

VERSION 1 – REVIEW

REVIEWER	Ali Jawad Barts Health NHS Trust QMUL
REVIEW RETURNED	25-Jun-2020

GENERAL COMMENTS	This is an interesting study and the results are presented clearly and concisely. But there are a number of limitations to this study in addition to those listed by the authors: 1. The authors selected patients with ILD from patients with 5 different underlying rheumatic diseases. Though ILD may be similar in those 5 different diseases, the underlying pathogenesis of these diseases is different. For example, lupus is an immune complex disease, dermatomyositis is mainly T cell mediated and the pathology of scleroderma is obliterative angiopathy with fibrosing process. RA is more complex and primary SS is mainly B cell mediated. This may have affected the findings. 2. The fact that patients with ILD were on higher doses of corticosteroids, methotrexate and and anti TNF is mostly due to the fact, these patients have more severe disease and more likely to get other complications of the underyling disease such as ILD. 3. The authors examined the the hourly levels of air pollutants 1 year before the index date of the development of ILD. We are not sure how early ILD develops and the assumption seems arbitrary. Perhaps the authors should have looked at a longer period of exposure, and possibly examined these factors even prior to the development of the underlying rheumatic disease.

REVIEWER	George Bertsias Faculty of Medicine, University of Crete, Greece
REVIEW RETURNED	05-Jul-2020

GENERAL COMMENTS	The Authors have used data from the Taiwan National Health Insurance Research Database to identify incident cases of a variety of CTDs and associated interstitial lung disease (ILD). By correlating these cases with the measurement levels of air pollutants during the year prior to diagnosis, they report on the inverse association between ozone levels and CTD-ILD. To

	overcome potential bias (especially confounding bias), they have matched cases with controls and have also adjusted for a variety of confounding factors. Still, one cannot entirely exclude the possibility that results are confounded by other clinical parameters. I have some comments listed below. 1. CTDs can cause a variety of lung disorders which I am unsure if they have been completely captured under the umbrella of "ILD" (and associated ICD codes). For instance, RA is known to cause "organising pneumonia" or similar radiologic presentations, which might have been missed. The authors need to elaborate and explain how they dealt with such cases. 2. Likewise, patients might experience respiratory symptoms due to other co-existing lung disorders (e.g. asthma, COPD) thus, undergoing work-up and identifying interstitial lung disease too. Concomitant lung disorders were not included in the model/matching algorithm, which could have biased the findings. 3. Not all patients with CTD have the same risk for ILD. For instance, RA patients who are positive for anti-CCP are at particularly high risk. In the same context, more severe/aggressive CTDs are more likely to involve lung. Therefore, it is not obvious whether exposure to ozone is a "marker" for more aggressive CTD (which then, is linked to ILD) and any causal inferences cannot be drawn. Have the authors tried to explore the associations within subgroups of CTDs such as anti-CCP +ve vs. -ve, anti-DNA +ve vs. -ve (for SLE), anti-Scl70 +ve vs. -ve (for SSc) etc?

REVIEWER	Tackseung Jun Kyung Hee University, South Korea
REVIEW RETURNED	23-Aug-2020

GENERAL COMMENTS	Referee report for Air pollutants and development of interstitial lung disease in patients with connective tissue disease: a population-based study 1 Summary of the paper This paper examines whether exposure to air pollution influences the development of interstitial lung disease (ILD) for patients with connective tissue disease (CTD). The authors use the observed patient data from the National Health Insurance program where details of the diagnosis, demographic information, and residence information are available. It is matched with the observations on air pollutants in the atmosphere to characterize exposure to air pollutants. The model of multivariate logistics regression is applied to the sample where the odds of developing ILD is estimated by regressing on exposure to various air pollutants and demographic variables and pre-conditions of diseases. One of the main findings is that exposure to O3 is inversely associated with a decreased risk of ILD in patients with rheumatoid arthritis (RA) and systemic sclerosis (SSc). 2 The main comments • As the authors pointed out, the association between the ILD-CTD incidence and exposure to O3 remains in the literature. The paper renders support for the inverse relationship. The paper points to the

	quenching effect of O3 for the possible explanation which is already discussed in the previous studies. Therefore I do not find any significant contribution of the paper to the literature. • The paper is hard to follow, as the crucial information about the data construction statistical analysis is missing. Please see the specific comments for details. • This investigation in the paper can be potentially interesting. However other than referring to a few studies, the authors only mentioned that there is a complicated relationship between exposure to air pollutants and ILD, but did not provide a detailed mechanism that they can be associated with each other. Some of these materials are found in the discussion section, which can be moved to the background section. This information may be crucial for understanding the literature in order to associate them with the results of the paper. 3 The specific comments • It is not easy to figure out how the data set of the regression analysis is arranged by reading the descriptions in the text. A summary statistics of the variables is not reported. It seems to be that the data set has a cross-sectional nature where each individual occupies a single observation in the sample. Or does it have a panel structure where an individual is followed over the course of time? • Related to the above issue, It is not clear how exposure to air pollutants was treated in the model. The authors mentioned “The hourly levels of air pollutants 1 year before index date,” but it is unclear what it represents. Does it mean that exposure to air pollutants exactly a year ago is matched with the date of CTD (or ILD) diagnosis? Do the authors use the mean (or some summary measure) of the air pollutants? • Regarding statistical analysis, the regression model of the paper, multivariate logistic regression, is not explicitly specified in the paper and so it is impossible to the validity of the results. Without the details of the regression specification, it is not clear whether the p-values are based on the robust standard errors. • The presented results are hard to follow. For example, the range and the p-values of the estimates are always shown in the main text of the paper, when they can be found in the tables. • Lines 249-251, It may be better to mention that the numbers in this paragraph are based on Table 2. • In Tables 2 to 5, there is no point in reporting the range of the estimated coefficients as long as the p-values are reported.

VERSION 1 – AUTHOR RESPONSE

Reviewers' Comments to Author:
Reviewer: 1
Reviewer Name: Ali Jawad
Institution and Country: Barts Health NHS Trust, QMUL, UK
Please state any competing interests or state 'None declared': None

This is an interesting study and the results are presented clearly and concisely.
But there are a number of limitations to this study in addition to those listed by the authors:
Q1. The authors selected patients with ILD from patients with 5 different underlying rheumatic diseases. Though ILD may be similar in those 5 different diseases, the underlying pathogenesis of these diseases is different. For example, lupus is an immune complex disease, dermatomyositis is mainly T cell mediated and the pathology of scleroderma is obliterative angiopathy with fibrosing process. RA is more complex and primary SS is mainly B cell mediated. This may have affected the findings.
Reply

- We thank you for this crucial point and have acknowledged this point as a limitation. We also increased the statement that patients with distinct CTDs may share similar pro-fibrotic pathway in the development of ILD. Please refer to the limitation section (page 20, line 336-338)
- Page 20, line 336-338

Third, varied mechanisms may underlie distinct CTDs; however, patients with distinct CTDs might have similar pro-fibrotic pathways in the development of ILD.[31]

Q2. The fact that patients with ILD were on higher doses of corticosteroids, methotrexate and anti-TNF is mostly due to the fact, these patients have more severe disease and more likely to get other complications of the underlying disease such as ILD.
Reply

- We totally agree with this point. Although disease activity is unavailable in NHIRD, the comprehensive record of CTD-associated should largely reflect the disease activity. We have listed this point as a limitation of the present study (page 20, line 333-336)
- Page 20, line 333-336

Second, the disease activity of CTD is not recorded, but we believe that we have adjusted for the essential CTD-associated medications, which were comprehensively in NHIRD. We think the adjustment of medications should largely reflect the disease activity.

Q3. The authors examined the hourly levels of air pollutants 1 year before the index date of the development of ILD. We are not sure how early ILD develops and the assumption seems arbitrary. Perhaps the authors should have looked at a longer period of exposure, andpossibly examined these factors even prior to the development of the underlying rheumatic disease.
Reply

- We are grateful for this insightful suggestion. One recently published study reported that the average delay from the onset of symptoms to the referral to interstitial lung disease (ILD) centre was 0.9 years and the overall diagnostic delay was 2.1 years [Hoyer et al., 2019, PMID, 31126287]. Therefore, we have conducted further analyses using a longer period (2-year) of air pollutant exposure, and the results were consistent with the finding in the present study using 1-year exposure to air pollutants. Discussion regarding this point is now added on page 20, line 340-343, and supplemental table 1.
- Page 20, line 340-343

Furthermore, we have conducted further analyses using a longer period (2-year) of air pollutant exposure, and the results were consistent with the finding in the present study using 1-year exposure to air pollutants (Supplemental table 1).

Reviewer: 2
Reviewer Name: George Bertsias
Institution and Country: Faculty of Medicine, University of Crete, Greece
Please state any competing interests or state 'None declared': None declared

The Authors have used data from the Taiwan National Health Insurance Research Database to identify incident cases of a variety of CTDs and associated interstitial lung disease (ILD). By correlating these cases with the measurement levels of air pollutants during the year prior to diagnosis, they report on the inverse association between ozone levels and CTD-ILD. To overcome potential bias (especially confounding bias), they have matched cases with controls and have also adjusted for a variety of confounding factors. Still, one cannot entirely exclude the possibility that results are confounded by other clinical parameters. I have some comments listed below.
Q1. CTDs can cause a variety of lung disorders which I am unsure if they have been completely captured under the umbrella of "ILD" (and associated ICD codes). For instance, RA is known to cause organising pneumonia or similar radiologic presentations, which might have been missed. The authors need to elaborate and explain how they dealt with such cases.
Reply

- We are grateful for this insightful suggestion and have included those with a diagnosis of organisingpneumonia (ICD: 516.36, n=10) for the analyses in the revised manuscript, and the results were consistent with the previous analyses.Given the existence of potential misclassificaion in NHIRD, we have acknowledged this point as a limitation, please refers to page 19-20, line 327-333.
- Page 19-20, line 327-333

Similarly, the accuracy of ILD in claim is also a concern. One recently published study aimed to validate claims-based algorithms for identification of ILD in patients with RA found that the accuracy of RA-ILD was high if the diagnosis was made by the specialist.[30]In the present study, we merely enrolled patients within the aforementioned catastrophic illness registry file. Therefore, the diagnoses of CTD and ILD were made by the rheumatologist, and the risk for misclassification should be at least partly mitigated.

Q2. Likewise, patients might experience respiratory symptoms due to other co-existing lung disorders (e.g. asthma, COPD) thus, undergoing work-up and identifying interstitial lung disease too. Concomitant lung disorders were not included in the model/matching algorithm, which could have biased the findings.
Reply

- We thank you for this comment. We have used Carlson comorbidity index (CCI) to adjust comorbidities, andpulmonary disease is one of the items among CCI. To further elaborate this concern, we have separated COPD and asthma from the CCI, and the data in the revised manuscript were consistent withhe finding in the previous manuscript.

Q3. Not all patients with CTD have the same risk for ILD. For instance, RA patients who are positive for anti-CCP are at particularly high risk. In the same context, more severe/aggressive CTDs are more likely to involve lung. Therefore, it is not obvious whether exposure to ozone is a "marker" for more aggressive CTD (which then, is linked to ILD) and any causal inferences cannot be drawn. Have the authors tried to explore the associations within subgroups of CTDs such as anti-CCP +ve vs. -ve, anti-DNA +ve vs. -ve (for SLE), anti-Scl70 +ve vs. -ve (for SSc) etc?
Reply

- We thank you for this comment and have listed the lack of laboratory data as the first limitation (page 19, line 322-326). Additionally, we also listed this limitation in the short bullet points of the present study (page 5, line 73-74).
- Page 19, line 322-326

First, the NHIRD cannot provide laboratory data including titers of autoantibody; however, the medication data are comprehensive. In addition, the diagnoses of SLE, RA and SS were validated by at least two experienced and qualified rheumatologists by reviewing patients' medical charts, laboratory findings and images to issue a catastrophic illness certificate.

- Page 5, line 73-74

Strengths and limitations of this study

4.Given the nature of the secondary data, the analysis misses some crucial variables, such as the disease activity and laboratory data.

Reviewer: 3

Reviewer Name: Tackseung Jun

Institution and Country: Kyung Hee University, South Korea

Please state any competing interests or state 'None declared': None declared

Please find the attached report.

Referee report for Air pollutants and development of interstitial lung disease in patients with connective tissue disease: a population-based study

Q1. Summary of the paper

This paper examines whether exposure to air pollution influences the development of interstitial lung disease (ILD) for patients with connective tissue disease (CTD). The authors use the observed patient data from the National Health Insurance program where details of the diagnosis, demographic information, and residence information are available. It is matched with the observations on air pollutants in the atmosphere to characterise exposure to air pollutants. The model of multivariate logistics regression is applied to the sample where the odds of developing ILD is estimated by regressing on exposure to various air pollutants and demographic variables and pre-conditions of diseases. One of the main findings is that exposure to O3 is inversely associated with a decreased risk of ILD in patients with rheumatoid arthritis (RA) and systemic sclerosis (SSc).

Reply

- We are grateful for the thoughtful reading of the reviewer and the following insightful comments.

Q2. The main comments

• As the authors pointed out, the association between the ILD-CTD incidence and exposure to O3 remains in the literature. The paper renders support for the inverse relationship. The paper points to the quenching effect of O3 for the possible explanation which is already discussed in the previous studies. Therefore I do not find any significant contribution of the paper to the literature.

Reply

- We have revised the introduction to explicit the rationale and niche of the present study. In addition to quenching effect, we also added evidence that O3 may exert the protective effect on incident ILD through modulating Th1/Th2 balance. Indeed, most studies have shown that O3 was positively associated with exacerbation of ILD, and few studies suggested the potential inverse correlation between O3 and incident ILD. Notably, no studies have been performed to explore the association between O3 and ILD in patients with CTDs. Therefore, the present population-based study focusing on patients with CTD provides crucial evidence for this niche.

Q3. The paper is hard to follow, as the crucial information about the data construction statistical analysisis missing. Please see the specific comments for details.
Reply

- We thank you for this comment and have added description regarding the regression model.In brief, we specified the data source, identification of CTD from the population, identification of ILD cases from the CTD cohorts, selection of matched non-ILD controls from the CTD cohort, measurement of exposure to air pollutants, potential confounders and statistical analyses, and the aforementioned descriptions should have given the readers a clear insight of study-design and analyses of the present study. Moreover, we have increased description with regards to potential confounders (page 10, line 163-166) and statistical analyses (page 12, line 196-197).
- Page 10, line 163-166 (potential confounders section)

The factors that may affect the association between exposure to air pollutants and incident ILD were taken into account as the confounder in the regression to estimate the impact of air pollutant on incident ILD in patients with CTD. Potential confounders that were adjusted for in the multivariable logistic regression model included

- Page 12, line 196-197 (statistical analyses section)

Variables were considered as candidates for inclusion in the multivariable model if the associated univariate p-value was lower than 0.20.19

Q4. This investigation in the paper can be potentially interesting. However other than referring to a few studies, the authors only mentioned that there is a complicated relationship between exposure to air pollutants and ILD, but did not provide a detailed mechanism that they can be associated with each other. Some of these materials are found in the discussion section, which can be moved to the background section. This information may be crucial for understanding the literature in order to associate them with the results of the paper.
Reply

- We are truly grateful for this insightful suggestion and have moved the description regarding potential mechanisms from the discussion section to the background section. Given that the present work is an epidemiological study, we think that it would be imprudent to propose a detailed mechanism.

Q5. The specific comments
• It is not easy to figure out how the data set of the regression analysis is arranged by reading the descriptions in the text. A summary statistics of the variables is not reported. It seems to be that the data set has a cross-sectional nature where each individual occupies a single observation in the sample. Or does it have a panel structure where an individual is followed over the course of time?
Reply

- We are truly grateful for this comment. We have added Fig. 1 to illustrate the time series of the study; therefore, the readers should clearly understand the panel structure of the present study.

Q6. Related to the above issue, It is not clear how exposure to air pollutants was treated in the model. The authors mentioned “The hourly levels of air pollutants 1 year before index date,” but it is unclear what it represents. Does it mean that exposure to air pollutants exactly a year ago is matched with the date of CTD (or ILD) diagnosis? Do the authors use the mean (or some summary measure) of the air pollutants?
Reply

- We thank you for this suggestion and have revised the manuscript to clearly point out that we used “mean” level or air pollutants (page 10, line 155).

Q7. Regarding statistical analysis, the regression model of the paper, multivariate logistic regression, is not explicitly specified in the paper and so it is impossible to the validity of the results. Without the

details of the regression specification, it is not clear whether the p-values are based on the robust standard errors.
Reply

- We thank you for this comment and have added a detailed description of the regression model. In the present study, we used p=0.2 to select potential confounders in the regression model. We have added one widely cited reference that elaborated the issue of selecting variables for the logistic regression model in the medical research.
- Page 12, line 196-202

Variables were considered as candidates for inclusion in the multivariable model if the associated univariate p value was lower than 0.20.[19] The association between the risk of ILD development and the exposure to air pollutants was examined using a multivariable conditional logistic regression analysis after adjusting for age, gender, CCI, urbanisation level, level of payroll-related insured amount and medications for CTD and is represented as adjusted odds ratio (aOR) with 95% confidence intervals (CIs).

Q8. The presented results are hard to follow. For example, the range and the p-values of the estimates are always shown in the main text of the paper, when they can be found in the tables.
Reply

- We agreed with you that p-value is redundant and have removed p-value from the manuscript (p-value is kept in the description of data in Table 1).

Q9. Lines 249-251, It may be better to mention that the numbers in this paragraph are based on Table 2.
Reply

- We thank for this suggestion and have added “Table 2” to avoid any confusion of the readers (Page 15, line 252).

Q10. In Tables 2 to 5, there is no point in reporting the range of the estimated coefficients as long as the p-values are reported.
Reply

- We do agree this remind and have removed p-value from Table 2 to 5.

VERSION 2 – REVIEW

REVIEWER	Ali Jawad Barts Health NHS Trust QMUL UK
REVIEW RETURNED	30-Sep-2020

GENERAL COMMENTS	Minor points Page 44 line 90: ILC should be clarified: Innate lymphoid cells. Page 58 line 336: Change 'think' to 'believe'.

REVIEWER	George Bertsias University of Crete Medical School
REVIEW RETURNED	09-Oct-2020

GENERAL COMMENTS	The Authors have addressed the points raised by the Referees and the manuscript has been improved

REVIEWER	Tackseung Jun

	Kyung Hee University, South Korea
REVIEW RETURNED	12-Oct-2020

GENERAL COMMENTS	Referee report for the 1st revised version of Air pollutants and development of interstitial lung disease in patients with connective tissue disease: a population-based case-control study in Taiwan 1 The main comments on the revised paper The comments I have made in the initial submission — detailed description of the statistical methods and data, and relevance of the paper to the literature — are mostly considered and incorporated into the revised paper by the authors. The paper is now self-contained, and readers should be able to follow the method, and results. I have a few minor comments below to make the paper more visible, and clear. 2 The specific comments • In Tables, the report of statistical significance can be expressed with *. For example, the coefficient that is 99% significant can be denoted by superscript of ***, the coefficient of 95% by superscript of ** and the coefficient of 90% by superscript of **. In this way, one can identify the level of significance. The current presentation only gives the 95% CI. • p.10 line 155: This is related to the point I raised in comment 6. The authors mentioned that the mean level of air pollutants was used to represent the air quality. It still does not answer my question: Does it mean that the mean level of air pollutants exactly a year ago is matched with patient data? I am not sure yet about how the air quality measure is linked to patient data. • p.10 line 159-160: "The ambient air pollutant concentrations at each residential location were estimated using a spatio-temporal model built via a deep-learning approach." The authors need to explain briefly why this approach is appropriate for the analysis.

VERSION 2 – AUTHOR RESPONSE

Reviewer-1

Name: Ali Jawad

Institution and Country: Barts Health NHS Trust

QMUL

UK

Please leave your comments for the authors below

Minor points

Q1. Page 6, line 90: ILC should be clarified: Innate lymphoid cells.

Reply

- We have added the abbreviation of ILC, please refers to page 6, line 90.

Q2. Page 20, line 336: Change 'think' to 'believe'.

Reply

- We have substituted “think” by “believe”. Please refer to page 21, line 342.

Reviewer-2

Reviewer Name: George Bertsias

Institution and Country: University of Crete Medical School

Please leave your comments for the authors below

The Authors have addressed the points raised by the Referees and the manuscript has been improved

Reply

- We sincerely thank the reviewer for the insightful suggestion in the revision process.

Reviewer-3

Reviewer Name: Tackseung Jun

Institution and Country: Kyung Hee University, South Korea

1 The main comments on the revised paper

The comments I have made in the initial submission — detailed description of the statistical methods and data, and relevance of the paper to the literature — are mostly considered and incorporated into the revised paper by the authors. The paper is now self-contained, and readers should be able to follow the method, and results. I have a few minor comments below to make the paper more visible, and clear.

Reply

- We are grateful for the through readings and insightful suggestions in the revision process and believe the current manuscript is clear for the reader.

2 The specific comments

Q1. In Tables, the report of statistical significance can be expressed with *. For example, the coefficient that is 99% significant can be denoted by superscript of ***, the coefficient of 95% by superscript of ** and the coefficient of 90% by superscript of **. In this way, one can identify the level of significance. The current presentation only gives the 95% CI.

Reply

- We thank the reviewer for this suggestion and have added “* p<0.05” and “** p<0.005” in the tables.

Q2. p.10 line 155: This is related to the point I raised in comment 6. The authors mentioned that the mean level of air pollutants was used to represent the air quality. It still does not answer my question: Does it mean that the mean level of air pollutants exactly a year ago is matched with patient data? I am not sure yet about how the air quality measure is linked to patient data.

Reply

- The Taiwan Environmental Protection Agency provides hourly levels of air pollutants measured by 60 air quality monitoring stations across Taiwan. We used the aforementioned raw data and calculated the mean level of air pollutants one year and two years before the index-date. To avoid the redundancy in the manuscript, we presented the data regarding 2-year exposure in the supplemental table. The descriptions regarding this point are now added on page 10, line 154-158 and page 21, line 347-350.

- Page 10, line 254-158

The hourly levels of air pollutants across from 60 air quality monitoring stations were used to calculate the mean level of exposed air pollutants, including particulate matter <2.5 μm in size (PM2.5), particulate matter <10 μm in size (PM10), nitrogen dioxide (NO2), carbon monoxide (CO), sulphur dioxide (SO2) and ozone (O3), one year prior to the index date

- Page 21, line 347-350

Furthermore, we have conducted further analyses using a longer period (2-year) of air pollutant exposure, and the results were consistent with the finding in the present study using 1-year exposure to air pollutants (Supplemental Table 1).

Q3. p.10 line 159-160: "The ambient air pollutant concentrations at each residential location were estimated using a spatio-temporal model built via a deep-learning approach." The authors need to explain briefly why this approach is appropriate for the analysis.

Reply

- We have added description regarding the applied deep-learning approach, please refers to page 10, line 160-163.

- Page 10, line 160-163

In brief, we used graph convolutional neural network to estimate the level of air pollutants at each residential locations, and the ambient level of air pollutants at 374 residential locations across Taiwan was estimated based on the data of three air quality monitoring stations near the location.