Factor analysis of household factors: are they associated with respiratory conditions in Chinese children?

Zhengmin Qian,1,2 Junfeng (Jim) Zhang,2 Leo R Korn,2,3 Fusheng Wei4 and Robert S Chapman5

Accepted 26 June 2003

Background We explored methods to develop uncorrelated variables for epidemiological analysis models. They were used to examine associations between respiratory health outcomes and multiple household risk factors.

Methods We analysed data collected in the Four Chinese Cities Study (FCCS) to examine health effects on prevalence rates of respiratory symptoms and illnesses in 7058 school children living in the four Chinese cities: Lanzhou, Chongqing, Wuhan, and Guangzhou. We used factor analysis approaches to reduce the number of the children’s lifestyle/household variables and to develop new uncorrelated ‘factor’ variables. We used unconditional logistic regression models to examine associations between the factor variables and the respiratory health outcomes, while controlling for other covariates.

Results Five factor variables were derived from 21 original variables: heating coal smoke, cooking coal smoke, socioeconomic status, ventilation, and environmental tobacco smoke (ETS) and parental asthma. We found that higher exposure to heating coal smoke was associated with higher reporting of cough with phlegm, wheeze, and asthma. Cooking coal smoke was not associated with any of the outcomes. Lower socioeconomic status was associated with lower reporting of persistent cough and bronchitis. Higher household ventilation was associated with lower reporting of persistent cough, persistent phlegm, cough with phlegm, bronchitis, and wheeze. Higher exposure to ETS and the presence of parental asthma were associated with higher reporting of persistent cough, persistent phlegm, cough with phlegm, bronchitis, wheeze, and asthma.

Conclusions Our study suggests that independent respiratory effects of exposure to indoor air pollution, heating coal smoke, and ETS may exist for the studied children.

Keywords Indoor air pollution, respiratory health effect, exposure assessment, China

There is growing evidence that indoor air pollution is a risk factor for the development of respiratory symptoms and illnesses in children.1–4 In the current literature of air pollution epidemiology, household gas combustion has often been considered a major source of indoor air pollution.5–8 The reason is that coal and other ‘dirty’ solid fuels are rarely used in Western countries where most available studies have been conducted. In many parts of the world, including China, another major

1 Department of Health Evaluation Sciences, Pennsylvania State University College of Medicine, 600 Centerview Drive, PO Box 855, Hershey, PA 17033, USA.
2 Environmental and Occupational Health Sciences Institute, University of Medicine and Dentistry of New Jersey and Rutgers University, 170 Frelinghuysen Road, Piscataway, NJ 08854, USA.
3 New Jersey Department of Environmental Protection, Division of Science, Research and Technology and the Division of Biometrics, UMDNJ-SPH, 401 East State Street, Trenton, NJ 08625–0409, USA.
4 China National Environmental Monitoring Center, Beishuangdonglu, Chaoyang District, Beijing 100029, China.
5 Formerly with National Center for Environmental Assessment, US Environmental Protection Agency, Mail Drop B243–01, Research Triangle Park, NC 27711, USA.
Correspondence: Dr Z Qian, 600 Centerview Drive, PO Box 855, Hershey, PA 17033, USA. E-mail: zqian@psu.edu
source of indoor air pollution is household coal combustion, used for heating and cooking.\textsuperscript{9,10} For example, in the four Chinese cities of Chongqing, Guangzhou, Lanzhou, and Wuhan, approximately 51–71% of households used coal for heating or cooking.\textsuperscript{11,12} Burning coal generally produces more air pollutants, in greater amounts, than burning liquid fuels or gas.\textsuperscript{13} However, few studies have systematically examined the health effects of household coal combustion. The results of the health effects from a limited number of published studies are somewhat inconsistent.\textsuperscript{14–16} This inconsistency may partly result from uncertainties associated with indoor exposure characterization and classification.

Most analyses of health effects of indoor air pollution exposure have relied on the data collected through questionnaire surveys. In epidemiological settings, a questionnaire is usually designed to ask a large set of questions on the subjects’ attributes with the purpose of obtaining enough information for the subsequent exposure assessment. However, sometimes the questions may not be the direct indicators of true exposure variables. The true exposure variables may be difficult or impossible to define, or may not be directly measurable.\textsuperscript{17–19} Sorting out useful information from the large amount of data collected is always a challenging task but is critical to obtain objective and unbiased exposure assessment in an epidemiological study. In the Four Chinese Cities Study (FCCS), we established a large dataset using a standardized questionnaire survey approach.\textsuperscript{16} This dataset contains a large amount of information on household air pollution sources (e.g. types of fuels and tobacco smoking), house characteristics that affect indoor air pollution levels (e.g. house type and ventilation conditions), and activities affecting pollution exposure (e.g. cooking location and cooking frequencies). It is of public health importance to ascertain the degree to which the known indoor air pollution sources (coal smoke targeted specifically) affect health and to identify potential risk factors from the large set of household variables collected from the questionnaire survey.

Factor analysis has been widely used to identify and summarize many inter-relationships that exist among individual variables. In a factor analysis, inter-correlated variables are combined into a smaller number of new variables (factors). This may enable us to simplify the dataset and consequently gain insights about underlying risk factors and true exposures that are linked to adverse health effects.\textsuperscript{20} In this study, we hypothesize that exposure to household coal combustion and environmental tobacco smoke (ETS) is positively associated with high prevalence of respiratory conditions in children. We extracted a set of uncorrelated (orthogonal) household risk factors from a large set of original variables obtained through the questionnaire survey by using a factor analysis method. We then used the factor scores of the extracted factors as independent variables in subsequent logistic regression models to examine associations between exposure and health outcomes. We expected the use of factor-score based variables to minimize the multicollinearity problem that is present in conventional regression analyses and to limit uncertainties associated with indoor exposure assessment.\textsuperscript{21,22}

Methods

Subjects

The study subjects have been described in the previous papers.\textsuperscript{23,24} Briefly, the FCCS included four geographically distinct large cities in China. In each city, two districts were selected with one being an inner city district of relatively high ambient air pollution levels and the other being a suburban district of relatively low ambient air pollution levels. In each district, one primary school was chosen for the study with the exception of the Wuhan suburban district. Two primary schools were chosen in this district to obtain a sample size comparable to those in other districts. A total of 7817 questionnaires were distributed to all the students enrolled in the selected schools, and 7754 completed questionnaires were collected from 1993 to 1996. Some of the returned questionnaires were excluded from further analyses based on the following considerations: (1) if a child had resided in his/her community for less than 2 years; (2) if a child’s household used wood as the main heating and cooking fuel. (The percentage rates of wood used for heating and cooking were small, <5%, in seven of the eight districts. The inclusion of wood-using households may complicate the investigation of coal smoke effects.); and (3) if there were unanswered or ambiguously answered key questions that could not be clarified. After these screenings had been completed, a total of 7058 children remained in the subject pool (Table 1).

Table 1 Response rates and numbers of questionnaires usable for analysis in the eight districts, 1993–1996

<table>
<thead>
<tr>
<th>Residence period</th>
<th>2 years or more</th>
<th>&lt;2 years</th>
<th>Wood used as fuel</th>
<th>Insufficient data</th>
<th>Insufficient data</th>
<th>Not usable for analysis</th>
<th>Usable for further analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>No. of questionnaires</td>
<td>920</td>
<td>99.3</td>
<td>1382</td>
<td>98.8</td>
<td>1924</td>
<td>99.3</td>
<td>593</td>
</tr>
<tr>
<td>Questionnaires returned</td>
<td>810</td>
<td>99.6</td>
<td>1328</td>
<td>98.8</td>
<td>1924</td>
<td>99.3</td>
<td>593</td>
</tr>
<tr>
<td>Usable for analysis</td>
<td>760</td>
<td>94.1</td>
<td>1280</td>
<td>95.7</td>
<td>1824</td>
<td>96.9</td>
<td>583</td>
</tr>
</tbody>
</table>
were potentially associated with children's respiratory health
(whether the associations were indirectly related to the purpose
information (e.g. children's school grades and other non-
the following considerations: (1) if a variable only provided
These variables consisted of 222 fields in SAS database format.
ponded to individual questions in the survey questionnaire.
original variables in the study database. These variables corres-
unconditional logistic regression analysis. We first screened all
We conducted data analysis at two levels, factor analysis, and
Statistical analysis
We conducted data analysis at two levels, factor analysis, and
unconditional logistic regression analysis. We first screened all
original variables in the study database. These variables corres-
pended to individual questions in the survey questionnaire.
These variables consisted of 222 fields in SAS database format.
We excluded some variables from the factor analyses based on
the following considerations: (1) if a variable only provided
information (e.g. children’s school grades and other non-
household variables) that was not directly related to the purpose
of the current analysis, i.e. identifying household factors that
were potentially associated with children's respiratory health
outcomes. (After this step, 58 fields remained); and (2) if a
variable had close to a uniform response across all subjects and
it was less likely to have an association with the health out-
comes of concern. After the above two screenings, we obtained
51 fields that were recoded and constructed into either ordinal
variables or indicator variables. This resulted in the formation of
21 re-organized variables that were used in the factor analysis.25

We used the SAS Factor procedure and report the results
from the principal components analysis with varimax rotations
here. We chose variables with factor loadings >0.4 for inter-
pretation in the study.20,22 The names of the extracted factors
were based on their high loading variables (>0.4). In each un-
conditional logistic regression model,26 the dependent variable
was one of the respiratory health outcomes of concern (persistent
cough, persistent phlegm, cough and phlegm, wheeze, bronchitis,
or asthma). The independent variables were the factor scores of
the factors obtained from the factor analysis, along with age,
gender, and the district dummy variables.

Results
Table 2 summarizes the high loading variables on the five
factors generated from the factor analysis. The naming of factors
was based on the high loading variables in each of the five
factors. Factor 1 was highly loaded with household heating
in winter, coal used for heating, coal stoves used for heating,
and chimney used for heating stoves. Given the nature of the
loading variables in this factor, we named this factor ‘heating
cal smoke’. Factor 2 was highly loaded with coal used for
cooking, coal stoves used for cooking, outside cooking, apart-
ment, and one-story house. We named this factor ‘cooking coal
smoke’. Factor 3, called ‘socioeconomic status’, was highly loaded
with mother’s education level, mother’s occupation, father’s

Table 2 summary of the principal components analysis, Varimax
rotation

<table>
<thead>
<tr>
<th>Factor names and high-loading variables</th>
<th>Factor loadings</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Factor 1: Heating coal smoke</strong></td>
<td></td>
</tr>
<tr>
<td>Household heating in winter</td>
<td>0.88</td>
</tr>
<tr>
<td>Coal used for heating</td>
<td>0.88</td>
</tr>
<tr>
<td>Coal stoves used for heating</td>
<td>0.89</td>
</tr>
<tr>
<td>Chimney used for heating stoves</td>
<td>−0.42</td>
</tr>
<tr>
<td><strong>Factor 2: Cooking coal smoke</strong></td>
<td></td>
</tr>
<tr>
<td>Coal used for cooking</td>
<td>0.70</td>
</tr>
<tr>
<td>Coal stoves used for cooking</td>
<td>0.64</td>
</tr>
<tr>
<td>Outside cooking</td>
<td>−0.60</td>
</tr>
<tr>
<td>Apartment</td>
<td>−0.72</td>
</tr>
<tr>
<td>One-storey house</td>
<td>0.64</td>
</tr>
<tr>
<td><strong>Factor 3: Socioeconomic status</strong></td>
<td></td>
</tr>
<tr>
<td>Mother's education level</td>
<td>0.77</td>
</tr>
<tr>
<td>Mother's occupation</td>
<td>−0.76</td>
</tr>
<tr>
<td>Father's education level</td>
<td>0.77</td>
</tr>
<tr>
<td>Father's occupation</td>
<td>−0.76</td>
</tr>
<tr>
<td><strong>Factor 4: Ventilation</strong></td>
<td></td>
</tr>
<tr>
<td>Home ventilation apparatus</td>
<td>−0.52</td>
</tr>
<tr>
<td>Room smokiness degree during cooking</td>
<td>0.68</td>
</tr>
<tr>
<td><strong>Factor 5: ETS² and parental asthma</strong></td>
<td></td>
</tr>
<tr>
<td>Father's smoking status</td>
<td>0.45</td>
</tr>
<tr>
<td>Other household smokers</td>
<td>0.55</td>
</tr>
<tr>
<td>Parents’ asthma</td>
<td>0.52</td>
</tr>
</tbody>
</table>

² Environmental tobacco smoke.
education level, and father’s occupation. This factor indicated that parents’ education level was inversely associated with employment as a manual labourer. We identified ‘ventilation’ as Factor 4. It was loaded negatively on the variable home ventilation apparatus, but positively on the variable degree of smokiness in the room during cooking. This result indicated that ventilation apparatus (chimney, hood, or exhaust fan) was inversely associated with smokiness during cooking. Factor 5, ‘ETS and parental asthma’, was identified based on the three positively loaded variables of fathers’ smoking status, other household smokers, and parents’ asthma (Table 2).

Table 3 summarizes the customized odds ratios (OR) and 95% CI for children’s respiratory symptoms and illnesses. In the unconditional logistic regression model, the slope coefficient βj associated with an explanatory variable Xj represents the change in log odds for an increase of one unit in Xj. The odds ratio (OR, \( e^{\beta} \)) is the ratio of odds for a one-unit change in Xj. In the current analysis, a change of one unit in the factor scores may be too small to be considered important while a change of the difference between the third quartile (75%) and the first quartile (25%) may be more meaningful. We, therefore, report the customized OR in this paper with the purpose of presenting the OR for a Q-spread (25–75%) change in the factor scores.27–28

The children’s exposure to the heating coal smoke factor was significantly associated with cough with phlegm (OR = 1.29; 95% CI: 1.11, 1.50), wheeze (OR = 1.22; 95% CI: 1.02, 1.45), and asthma (OR = 1.52; 95% CI: 1.06, 2.15). The cooking coal smoke factor was associated with none of the six health outcomes. The lower socioeconomic status factor was significantly associated with lower reporting of persistent cough (OR = 0.81; 95% CI: 0.70, 0.94) and bronchitis (OR = 0.70; 95% CI: 0.63, 0.78). The higher household ventilation factor was significantly associated with lower reporting of persistent cough (OR = 0.73; 95% CI: 0.65, 0.83), persistent phlegm (OR = 0.78; 95% CI: 0.66, 0.93), cough with phlegm (OR = 0.82; 95% CI: 0.74, 0.89), bronchitis (OR = 0.85; 95% CI: 0.78, 0.93), and wheeze (OR = 0.80; 95% CI: 0.72, 0.89), but not asthma. The ETS and parental asthma factor was significantly associated with all of the six health outcomes; persistent cough (OR = 1.50; 95% CI: 1.35, 1.67), persistent phlegm (OR = 1.66; 95% CI: 1.43, 1.94), cough with phlegm (OR = 1.47; 95% CI: 1.35, 1.60), bronchitis (OR = 1.34; 95% CI: 1.24, 1.45), wheeze (OR = 1.62; 95% CI: 1.47, 1.78), and asthma (OR = 1.47; 95% CI: 1.22, 1.78) (Table 3).

**Discussion**

The results from the present study support our hypothesis that household coal combustion is positively associated with high prevalence of respiratory symptoms and illnesses in Chinese children. The results further indicate that it is exposure to the heating coal smoke factor, but not the cooking coal smoke, that is significantly associated with cough with phlegm, wheeze, and asthma. The results also show that socioeconomic status, household ventilation, presence of parental asthma, and exposure to ETS may be important indicators of the respiratory health for the children under study.

Surprisingly, the results show that lower socioeconomic status was significantly associated with lower reporting of persistent cough and bronchitis in children. There are two possible reasons that may explain the results. In this study, low socioeconomic.
status means that the parents had low education levels and were manual labourers. There might be a difference in rates of reporting based on the parents’ education level. That is, the better-educated parents might be more likely to report symptoms. Another reason is that the current analysis used data collected in China from 1993 to 1996 when society was unstable with many reforms being implemented. Lower socioeconomic status households, as defined in the current analysis, might not represent a group with low incomes as it would in developed countries where the level of socioeconomic status is generally in agreement with household incomes. For example, at that time parents with low education levels might have higher incomes than those with high education, and manual labourers might also obtain higher salaries than non-manual labourers. Households with high incomes tended to use ‘cleaner’ but more expensive fuel at home, such as gas or electricity. The children living in these households were then exposed to less indoor air pollutants from household heating and cooking activities and had weaker health effects. Due to practical concerns, family income information was not collected in any of the participating households.

Another ‘surprising’ finding of the current analysis is the lack of associations between any health outcomes and cooking coal smoke exposure. This observation, however, is consistent with a finding from our previous analysis using a different statistical method.29 We speculate this is mainly because the children under study usually stayed away from their homes when cooking took place and, hence, avoided exposure to peak concentrations of cooking coal smoke. The cooking coal smoke variable constructed here, therefore, may not represent children’s true exposure to cooking coal smoke.

To our knowledge, few epidemiological studies on air pollution have examined associations between respiratory health outcomes and exposure by developing multiple exposure variables using factor scores.30,31 The large sample size in the present study provided us with a good opportunity to explore ‘true’ exposure for the children under study by comparing the clustering features of the exposure-related variables in terms of factor patterns. By doing so, we reduced 21 lifestyle/household variables to five potential risk factors relevant to the children’s exposure in the four study cities. These results of the variable reduction suggest that the children’s exposure was multifactorial.32 These multiple risk factors might act as true exposure factors or confounding factors associated with children’s respiratory health. First, at the variable level, multiple correlated variables were grouped to form a factor. These variables included not only pollution sources but also exposure-related variables such as house type. The factor cooking coal smoke, for example, was composed of five variables, coal used for cooking, coal stoves used for cooking, outside cooking, apartment, and one-story house. Second, at the factor level, the risk factors could be categorized into pollution sources (heating coal smoke, cooking coal smoke), combination of pollution source and related variables (ETS and parental asthma), socioeconomic status, and a factor that influences indoor pollution levels (household ventilation).

The current analysis used factor scores as new variables in the unconditional logistic regression models to study the associations between children’s respiratory health outcomes and factor scores. This strategy has three advantages. First, it minimizes the multicollinearity problem since the five developed factors are orthogonal and the correlated variables are blocked within their factors. Second, potential confounding factors are included in the regression model since the identification of the confounding factors in the analysis is not only based on judgement but also on the results from the factor analysis. Third, the roles of each input variable are included when modelling the health outcome in the logistic regression models because the factor scores represent weighted combinations of the subject’s scores on each of the input variables.21,27 This effort prevented or reduced the possible omission of some important variables in the models. Therefore, we consider this approach preferable for revealing the true exposure among a large set of exposure-related variables from an epidemiological questionnaire survey.

Despite the obvious benefits mentioned above for the factor analysis, it is acknowledged that factor analysis is a complicated sequence of procedures, involving a great deal of subjective judgement. For example, determining the number of factors and labelling them were subjective. The selection of the 21 variables to form the factors was also subjective and merely reflected our preconception as to which pollution-related variables were relevant to the health outcomes under study. In performing the factor analysis, we assumed that the variables collected (through the questionnaire survey) were not necessarily the ones that we were interested in. The procedure was asked to find a set of orthogonal factors that presumably reflect exposure better than original variables. Doing this, however, may hamper us from examining some originally collected variables of interest. For example, Factor 5 consisted of three original variables (fathers’ smoking status, other household smokers, and parental asthma). Knowing the health effect of this factor, as a whole, is perhaps less interesting and of less practical importance than knowing the effects of the individual variables.

To separate the ETS effect and the parental asthma effect from the overall effect of Factor 5, we performed a new set of logistic regression analyses using the original variables that formed the 5 factors (Table 2), along with age, gender, and the district dummy variables, as independent variables. The results show that parental asthma was significantly associated with

<table>
<thead>
<tr>
<th>Variable</th>
<th>Cough OR 95% CI</th>
<th>Phlegm OR 95% CI</th>
<th>Cough with phlegm OR 95% CI</th>
<th>Bronchitis OR 95% CI</th>
<th>Wheeze OR 95% CI</th>
<th>Asthma OR 95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td>ETS</td>
<td>1.15 0.93, −1.43</td>
<td>1.24 0.90, −1.70</td>
<td>1.32 1.11, −1.56</td>
<td>1.29 1.11, −1.49</td>
<td>1.21 0.99, −1.47</td>
<td>1.80 0.69, −1.46</td>
</tr>
<tr>
<td>Parental asthma</td>
<td>1.86 1.53, −2.26</td>
<td>2.61 2.02, −3.38</td>
<td>2.15 1.84, −2.53</td>
<td>1.58 1.36, −1.83</td>
<td>3.22 2.73, −3.81</td>
<td>2.80 2.04, −3.84</td>
</tr>
</tbody>
</table>

* a Environment tobacco smoke.

* b The model was adjusted for original variables in Table 2 plus age, gender, and the eight study districts.

* c ETS was defined as either a father was a smoker or a household member was a smoker.
all six health outcomes; and that ETS exposure was associated significantly with cough with phlegm and bronchitis (Table 4). The ETS effects and parental asthma effects observed in the current analysis are generally in agreement with findings from a previous analysis using different models. These results suggest that the overall effects of Factor 5 were largely driven by parental asthma. Reasons for the strong parental asthma effects on children’s respiratory health may include the following. Parental asthma may be linked to increased susceptibility in their children to environmental exposure through gene–environment interactions. The parents and their children might have been exposed to similar levels of indoor air pollution since they had lived in the same houses. Further studies are needed to better understand this question.

Acknowledgement

The authors express their appreciation for data collection by our colleagues at the four cities and co-operation of the subjects. The study was partially funded by the US Environmental Protection Agency (EPA, Cooperative Agreements CR826507 with the University of Medicine and Dentistry of New Jersey and CR822691 with the China National Environmental Monitoring Center). Dr Zhang’s research is also supported in part by an NIEHS center grant # ES05022–10 to the Environmental and Occupational Health Sciences Institute (EOHSI). The views expressed in this article are those of the authors and do not necessarily reflect the views or policies of either the US EPA or the Chinese Government.

References


